Classifying Mushrooms as Edible or Poisonous

Arya Bharath and James Wright

10/21/24

Table of Contents

Statement & Goal

The Mushrooms Dataset takes values of 20 unique attributes in order to predict whether or not any given mushroom will be poisonous (p) or edible (e). Our goal for this project is to build effective models using this dataset that will accurately predict the class variable.

We plan to use this dataset to classify mushrooms as either poisonous (p) or edible (e) based on the 20 outlined attributes. In the real world, this has several implications. If someone sees or interacts with a mushroom that they suspect may be poisonous, for example, how will they know? It would be best to know as soon as possible if they will have symptoms from touching or interacting with something poisonous, but most people will not be able to identify whether something is definitively poisonous or not based on looking at it alone. But if this is the case, what is the optimal method for classifying? It would be possible to create a very simple set of rules that humans can understand and use without following more advanced algorithms, like stating that very colorful mushrooms are poisonous, but this would lose a lot of precision that would result in worse results over time. As a result, it would be useful to use a machine learning model to classify instead.

Description of Dataset

Link to Dataset:
https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

Data Information, Dimensionality, & Instances
There were originally 4 datasets downloaded from the ML Repository at UCI. The downloads included two .txt files which shared information about each of the attributes and what type of data they were, as well as two .csv files which held all of the instances and values. We downloaded both primary and secondary mushroom datasets, but we chose to only work with one file for the rest of our project. The file that we used was:

secondary_data.csv

We chose to use this over the primary dataset as it contained more missing data as well as better attributes.

Our dataset contains 61,069 instances of mushrooms, with each mushroom being given a label of poisonous or edible. The dimensionality of this dataset is 20, with each attribute representing key features of the mushroom's physical appearance and anatomy.

Attributes
Before preprocessing, our dataset has 20 attributes, expressing both qualitative and quantitative data. In combination, these attributes are used in order to predict whether a certain mushroom is poisonous (p) or edible (e). (**class**: qualitative nominal, given as poisonous (p) or edible (e))

1. **cap-diameter**: quantitative continuous, given as a float in centimeters (cm)
2. **cap-shape**: qualitative nominal, given as bell (b), conical (c), convex (x), flat (f), sunken (s), spherical (p), or others (o)
3. **cap-surface**: qualitative nominal, given as fibrous (i), grooves (g), scaly (y), smooth (s), shiny (h), leathery (l), silky (k), sticky (t), wrinkled (w), or fleshy (e)
4. **cap-color**: qualitative nominal, given as brown (n), buff (b), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y), blue (l), orange (o), or black (k)
5. **does-bruise-bleed**: qualitative nominal, given as bruises-or-bleeding (t) or no (f)
6. **gill-attachment**: qualitative nominal, given as adnate (a), adnexed (x), decurrent (d), free (e), sinuate (s), pores (p), none (f), or unknown (?)
7. **gill-spacing**: quantitative continuous, given as a float in centimeters (cm)
8. **gill-color**: qualitative nominal, given as brown (n), buff (b), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y), blue (l), orange (o), black (k), or none (f)
9. **stem-height**: quantitative continuous, given as a float in centimeters (cm)

10. **stem-width**: quantitative continuous, given as a float in centimeters (cm)
11. **stem-root**: qualitative nominal, given as bulbous (b), swollen (s), club (c), cup (u), equal (e), rhizomorphs (z), or rooted (r)
12. **stem-surface**: qualitative nominal, given as fibrous (i), grooves (g), scaly (y), smooth (s), shiny (h), leathery (l), silky (k), sticky (t), wrinkled (w), fleshy (e), or none (f)
13. **stem-color**: qualitative nominal, given as brown (n), buff (b), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y), blue (l), orange (o), black (k), or none (f)
14. **veil-type**: qualitative nominal, given as partial (p) or universal (u)
15. **veil-color**: qualitative nominal, given as brown (n), buff (b), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y), blue (l), orange (o), black (k), or none (f)
16. **has-ring**: qualitative nominal, given as ring (t) or none (f)
17. **ring-type**: qualitative nominal, given as cobwebby (c), evanescent (e), flaring (r), grooved (g), large (l), pendant (p), sheathing (s), zone (z), scaly (y), movable (m), none (f), or unknown (?)
18. **spore-print-color**: qualitative nominal, given as brown (n), buff (b), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y), blue (l), orange (o), or black (k)
19. **habitat**: qualitative nominal, given as grasses (g), leaves (l), meadows (m), paths (p), heaths (h), urban (u), waste (w), or woods (d)
20. **season**: qualitative nominal, given as spring (s), summer (u), autumn (a), or winter (w)

Missing Data

After running the data set through a Python script that counted the number of missing values, we determined that our data is missing 307,463 values total. While this may seem like a lot, it is important to note that this is spread out over 61,069 instances, meaning that each instance will be missing, on average, 5 values. Because we have 20 attributes, this should not be too detrimental to our models, particularly after we preprocess our data to deal with missing values. It is also interesting to note that because we have so many instances, it may be possible to simply select values with only a small number missing values to solve the data problem.

Class Distribution

In the overall dataset, there are 33,888 instances that have been classified as poisonous (p) and 27,181 classified as edible (e). As a result, because these values are not equal, our dataset is not balanced, with 55.5% of our instances being deemed poisonous and approximately 44.5% being edible.

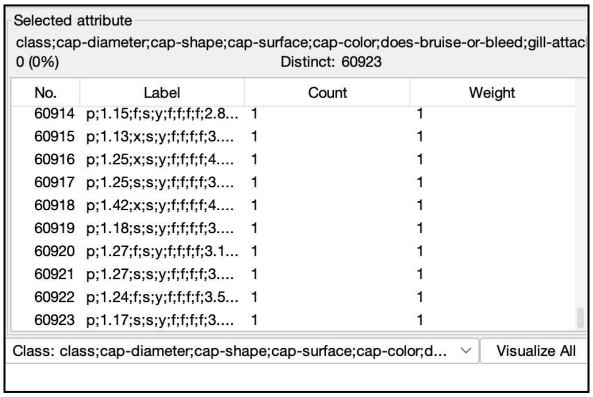
Uniform or Skewed

In terms of whether our data is uniform or skewed, it is hard to tell because the class is qualitative and nominal, meaning that there is no order. This would mean that I could arrange the data for the poisonous and edible mushrooms in any order, which would not create a uniform or skewed distribution.

Data Mining Procedure/Preprocessing

Preprocessing

So far, most of our work has gone into preprocessing our data. The first thing that we did was solve the problem of opening our data set in WEKA. Originally, our data looked like this:



Selected attribute

class;cap-diameter;cap-shape;cap-surface;cap-color;does-bruise-or-bleed;gill-attac

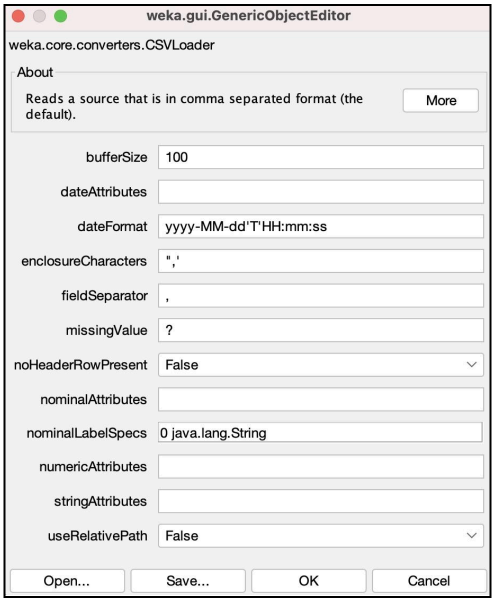
0 (0%) Distinct: 60923

No.	Label	Count	Weight
60914	p;1.15;f;s;y;f;f;f;f;2.8...	1	1
60915	p;1.13;x;s;y;f;f;f;f;3....	1	1
60916	p;1.25;x;s;y;f;f;f;f;4....	1	1
60917	p;1.25;s;s;y;f;f;f;f;3....	1	1
60918	p;1.42;x;s;y;f;f;f;f;4....	1	1
60919	p;1.18;s;s;y;f;f;f;f;3....	1	1
60920	p;1.27;f;s;y;f;f;f;f;3.1...	1	1
60921	p;1.27;s;s;y;f;f;f;f;3....	1	1
60922	p;1.24;f;s;y;f;f;f;f;3.5...	1	1
60923	p;1.17;s;s;y;f;f;f;f;3....	1	1

Class: class;cap-diameter;cap-shape;cap-surface;cap-color;d...

In the above screenshot, both the classes and instances appeared not to split apart. Even though there were approximately 60,000 instances in our data (as expected), each instance was given as a string of all of the values of the attributes combined into one. Similarly, as we can see in the middle of the screenshot, the attributes and class did not split up. In order to fix this problem, we identified that the issue was the formatting of the .csv file. Even though we expected commas to separate the values as is true of files of the type .csv, there were semicolons.

Our solution, as we posted in the Q1 Project Help document, was to edit WEKA's GenericObjectEditor menu to change the *fieldSeparator* from a comma to a semicolon.



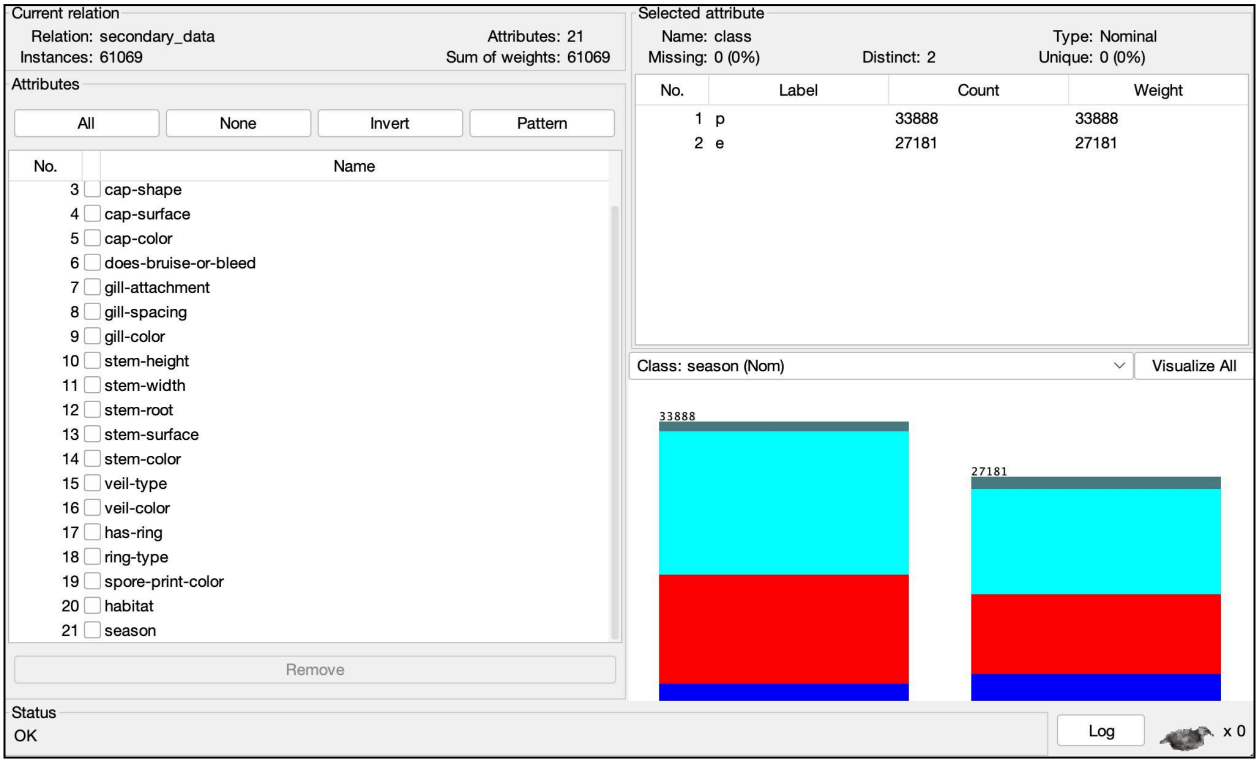
weka.core.converters.CSVLoader

About

Reads a source that is in comma separated format (the default).

bufferSize	100
dateAttributes	
dateFormat	yyyy-MM-dd'T'HH:mm:ss
enclosureCharacters	",'
fieldSeparator	,
missingValue	?
noHeaderRowPresent	False
nominalAttributes	
nominalLabelSpecs	0 java.lang.String
numericAttributes	
stringAttributes	
useRelativePath	False

This menu gives us a few pieces of information. On top of changing the *fieldSeparator* as previously mentioned, we also used the *missingValue* field to change our missing values to question marks.



Current relation

Relation: secondary_data Attributes: 21

Instances: 61069 Sum of weights: 61069

Attributes

No.	Name
3	□ cap-shape
4	□ cap-surface
5	□ cap-color
6	□ does-bruise-or-bleed
7	□ gill-attachment
8	□ gill-spacing
9	□ gill-color
10	□ stem-height
11	□ stem-width
12	□ stem-root
13	□ stem-surface
14	□ stem-color
15	□ veil-type
16	□ veil-color
17	□ has-ring
18	□ ring-type
19	□ spore-print-color
20	□ habitat
21	□ season

Selected attribute

Name: class Type: Nominal

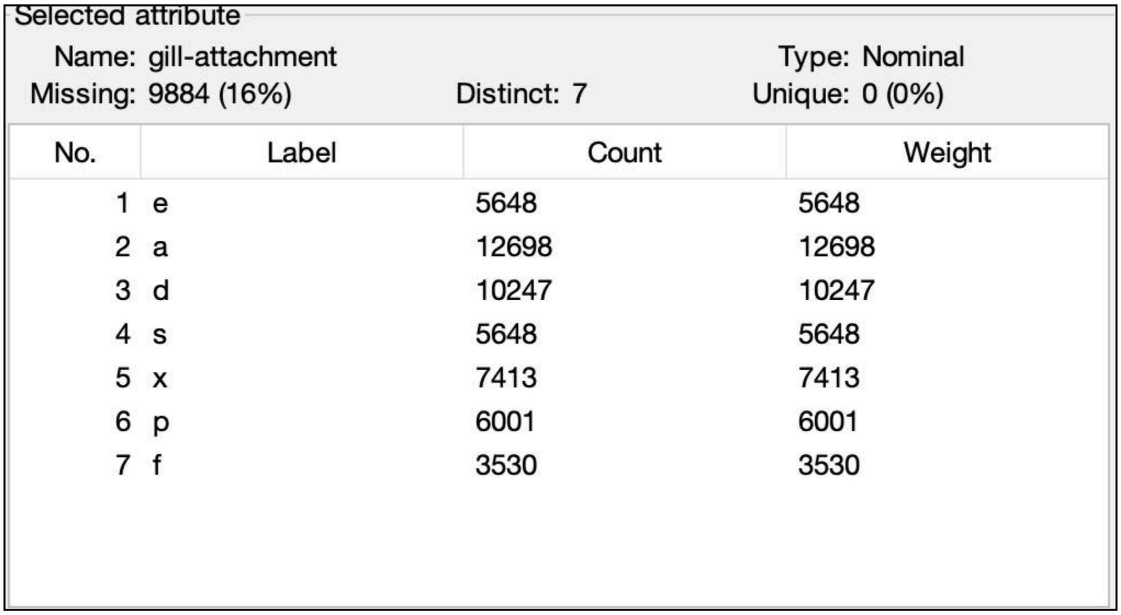
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	p	33888	33888
2	e	27181	27181

Status

OK

After opening our data, it now looks like this, which is much better. Additionally, as we can see from the screenshot below, WEKA has missing values identified when attributes are selected.



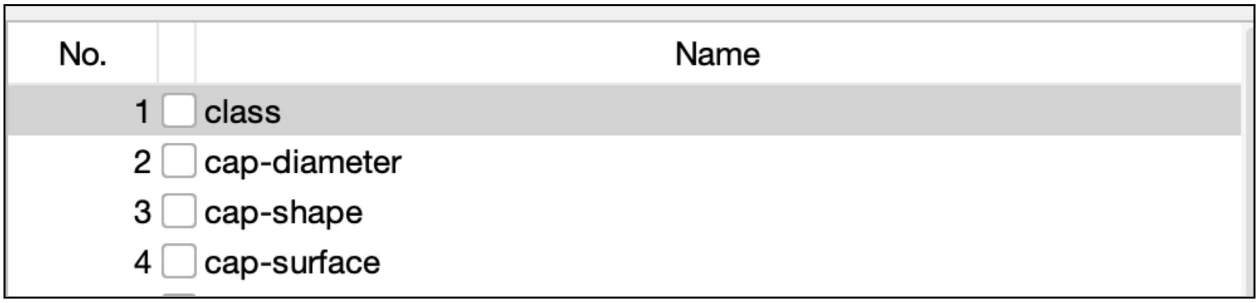
Selected attribute

Name: gill-attachment Type: Nominal
Missing: 9884 (16%) Distinct: 7 Unique: 0 (0%)

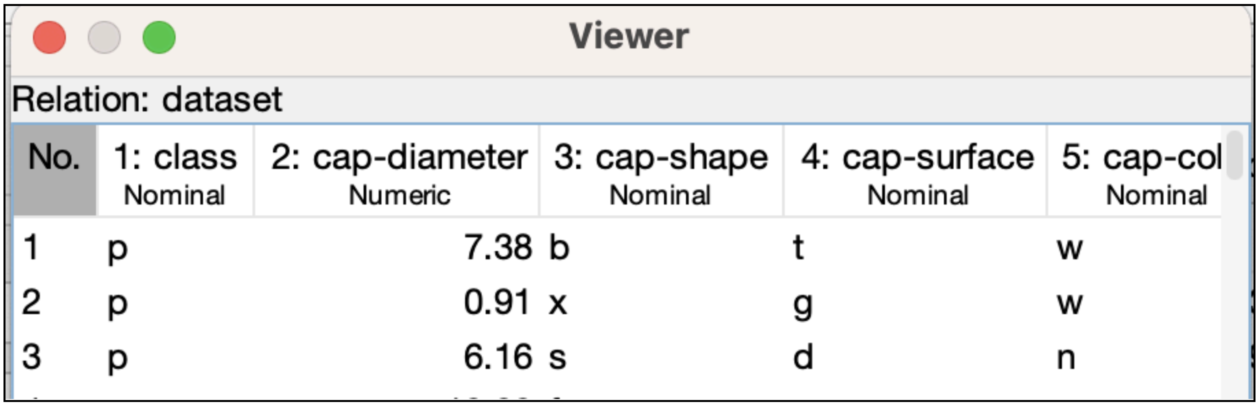
No.	Label	Count	Weight
1	e	5648	5648
2	a	12698	12698
3	d	10247	10247
4	s	5648	5648
5	x	7413	7413
6	p	6001	6001
7	f	3530	3530

However, in order to deal with missing values, we are trying two different approaches. Because of the nature of the missing values in our data, it is unclear if it's a good idea to use mean/mode to fill missing values, just because of the lack of information in certain columns. As a result, we are going to try 1) setting missing values as "?" unknown values and 2) setting them as mode/mean in WEKA. This means that we will have 30 datasets total, as there are 5 attribute selections and 3 sets per attribute selection (train, validation, test), and 2 approaches.

When opening our training dataset in WEKA, the class label appeared first, as an attribute.



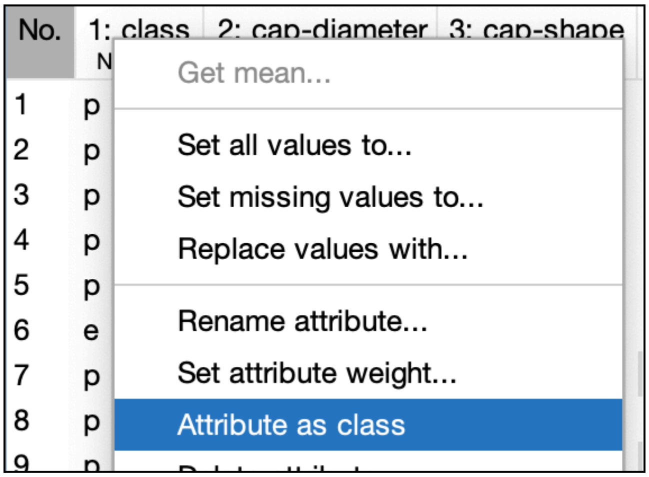
No.		Name
1	□	class
2	□	cap-diameter
3	□	cap-shape
4	□	cap-surface



Relation: dataset

No.	1: class Nominal	2: cap-diameter Numeric	3: cap-shape Nominal	4: cap-surface Nominal	5: cap-col Nominal
1	p	7.38	b	t	w
2	p	0.91	x	g	w
3	p	6.16	s	d	n

To make the class an actual class in WEKA instead of just an attribute, we opened the Viewer using Edit and right-clicked on the class, which then prompted us with the option to set the attribute as the class.



No.	1: class	2: cap-diameter	3: cap-shape
1	p		
2	p		
3	p		
4	p		
5	p		
6	e		
7	p		
8	p		

After clicking that, we then clicked “OK” on the Viewer to set that class attribute as the real class in WEKA. We also considered using normalization in our data. However, all of the quantitative values were constrained to the same ranges of numbers, so we decided that it was not necessary.

To split our dataset into training, validation, and testing sets, we used a Python script that employed pandas, sklearn, scipy.io, and the arff libraries:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from scipy.io import arff
import arff as liac_arff

data, meta = arff.loadarff('data.arff')
df = pd.DataFrame(data)
```

```
      def decode_bytes(df):
          for col in df.select_dtypes([object]):
              df[col] = df[col].apply(lambda x: x.decode('utf-8') if
isinstance(x, bytes) else x)
          return df

      df = decode_bytes(df)

      X = df.iloc[:, :-1]
      y = df.iloc[:, -1]

      X_train, X_temp, y_train, y_temp = train_test_split(X, y,
test_size=0.3, stratify=y, random_state=42)

      X_val, X_test, y_val, y_test = train_test_split(X_temp, y_temp,
test_size=0.5, stratify=y_temp, random_state=42)

      train_set = pd.concat([y_train, X_train], axis=1)
      val_set = pd.concat([y_val, X_val], axis=1)
      test_set = pd.concat([y_test, X_test], axis=1)

      def save_to_arff(df, filename, meta):
          arff_data = {
              'description': '',
              'relation': 'dataset',
              'attributes': [(name, list(df[name].unique())) if
df[name].dtype == 'object' else (name, 'REAL') for name in df.columns],
              'data': df.values.tolist()
          }
          with open(filename, 'w') as f:
              liac_arff.dump(arff_data, f)

      save_to_arff(train_set, 'train_set.arff', meta)
      save_to_arff(val_set, 'val_set.arff', meta)
      save_to_arff(test_set, 'test_set.arff', meta)
```

When using the train_test_split method from sklearn.model_selection, we used the stratify parameter to ensure that the training, validation, and testing datasets would all maintain the same proportions between class labels. For our data, the ratio between p (poisonous) to e (edible) instances was around 1.25, so all of the split datasets would also need to have this ratio. We used a 70:15:15 split between the training, validation, and testing datasets when splitting each dataset created from attribute selection. Because we created 10 different files from attribute selection (5 from every different evaluation and another 5 from filling in all missing values), 30 total files were created of training, validation, and testing datasets. The training, validation, and testing datasets always had 42672 instances, 9174 instances, and 9175 instances respectively.

Data Mining Tools & Algorithms/Attribute Selection

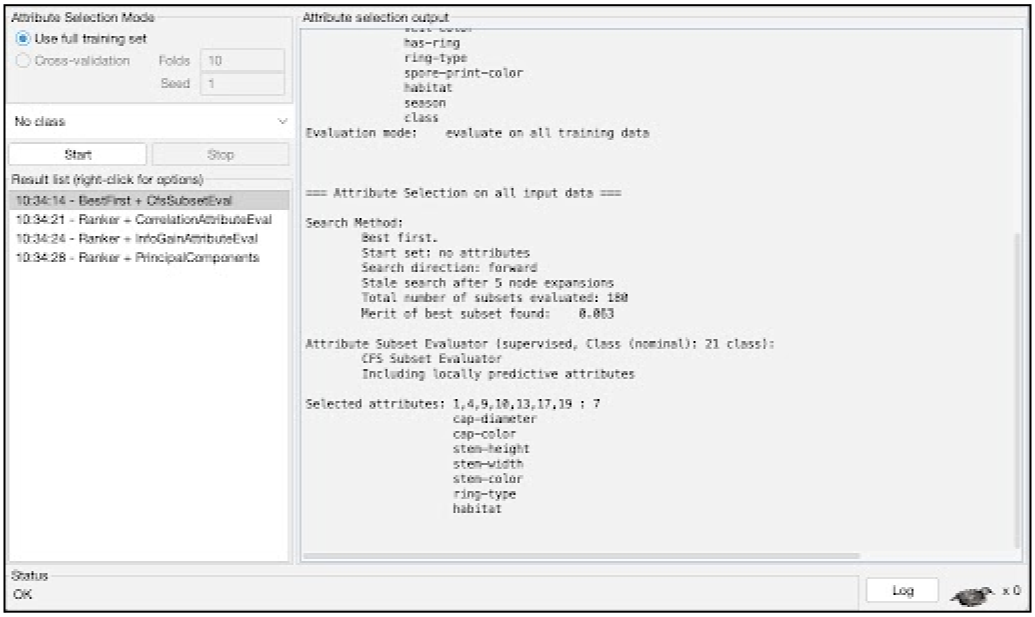
Attribute Selection

For this project, we needed to choose 5 different approaches to do feature selection. Ultimately, we decided on CfsSubsetEval, CorrelationAttributeEval, InfoGainAttributeEval, OneRAttributeEval, and our own intuitive selection. As mentioned previously, we want to compare several different ways of preprocessing missing values in order to see which one yields the best performances, so we ran each attribute selection algorithm for a dataset with each preprocessing method. In the future, we may also use Principal Component Analysis to reduce our dimensionality as well.

a) CfsSubsetEval

CfsSubsetEval uses a combination of the individual predicting power of each attribute and the redundancy of the predictions to select attributes.

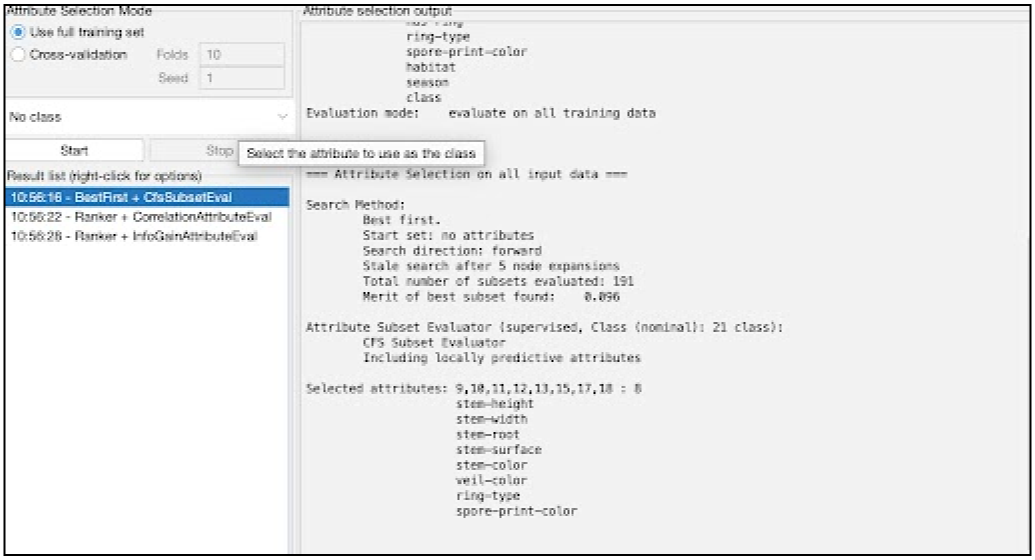
For our datasets with missing values as "?":



10:34:14 - BestFirst + CfsSubsetEval
10:34:21 - Ranker + CorrelationAttributeEval
10:34:24 - Ranker + InfoGainAttributeEval
10:34:28 - Ranker + PrincipalComponents

```
              has-ring
              ring-type
              spore-print-color
              habitat
              season
              class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 180
        Merit of best subset found:    0.063

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,4,9,10,13,17,19 : 7
                     cap-diameter
                     cap-color
                     stem-height
                     stem-width
                     stem-color
                     ring-type
                     habitat
```

For our datasets with missing values as mean/mode:

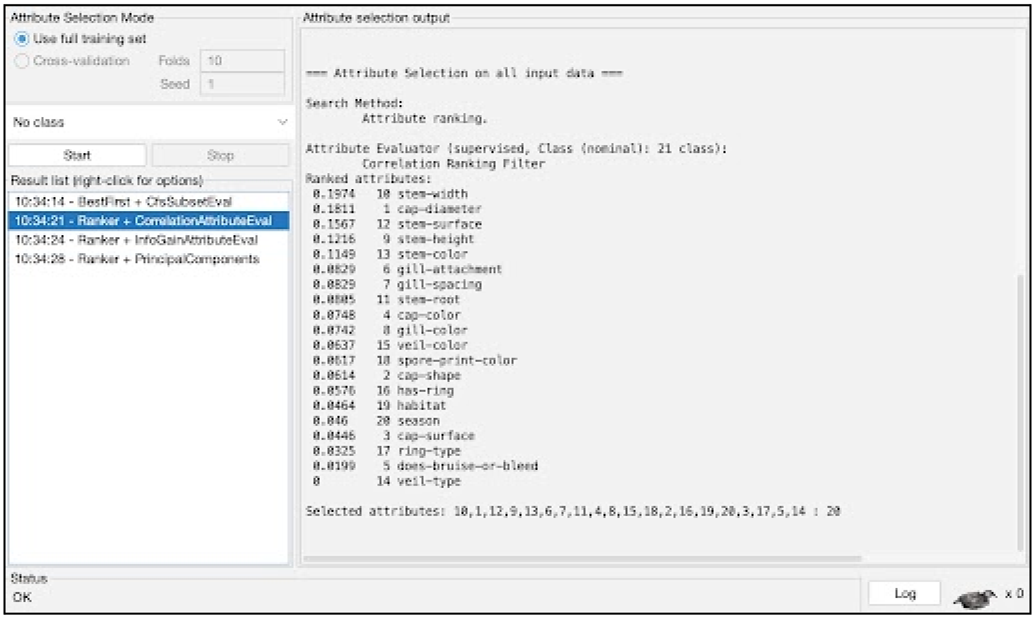

Attribute Selection Mode
● Use full training set
○ Cross-validation
Folds 10
Seed 1
No class
Start
Stop
Select the attribute to use as the class
Result list (right-click for options)
10:56:16 - BestFirst + CfsSubsetEval
10:56:22 - Ranker + CorrelationAttributeEval
10:56:28 - Ranker + InfoGainAttributeEval
Attribute selection output

```
              has-ring
              ring-type
              spore-print-color
              habitat
              season
              class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 191
        Merit of best subset found:    0.096

Attribute Subset Evaluator (supervised, Class (nominal): 21 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 9,10,11,12,13,15,17,18 : 8
                     stem-height
                     stem-width
                     stem-root
                     stem-surface
                     stem-color
                     veil-color
                     ring-type
                     spore-print-color
```

b) CorrelationAttributeEval

Orders attributes by using Pearson Correlation between each individual attribute and the class.

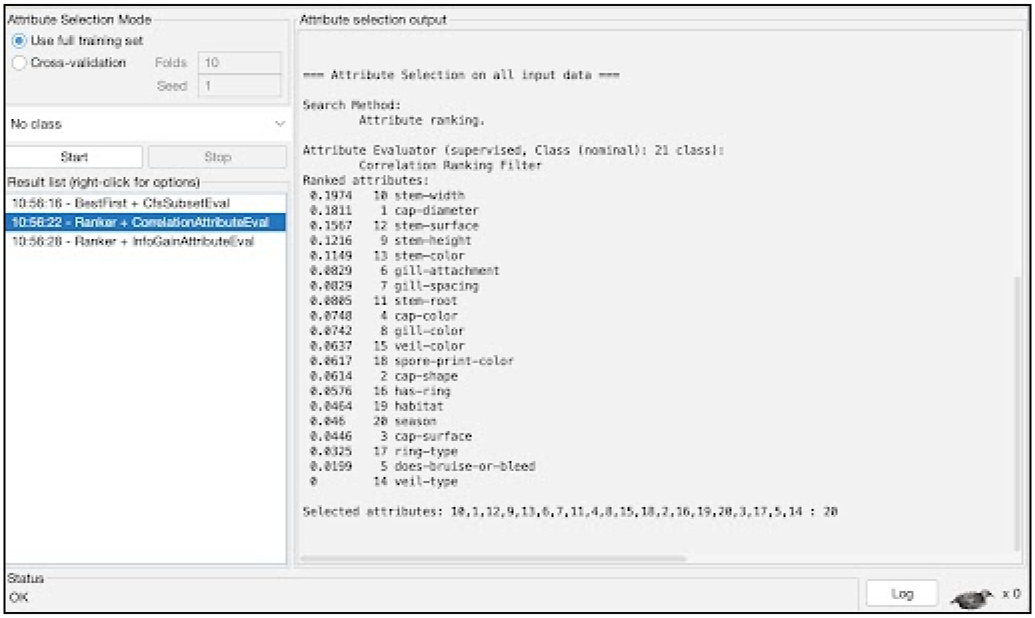
For our datasets with missing values as "?":


Attribute Selection Mode
● Use full training set
○ Cross-validation
Folds 10
Seed 1
No class
Start
Stop
Result list (right-click for options)
10:34:14 - BestFirst + CfsSubsetEval
10:34:21 - Ranker + CorrelationAttributeEval
10:34:24 - Ranker + InfoGainAttributeEval
10:34:28 - Ranker + PrincipalComponents
Attribute selection output
Status
OK
Log
x 0

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Correlation Ranking Filter
Ranked attributes:
 0.1974   10 stem-width
 0.1811    1 cap-diameter
 0.1567   12 stem-surface
 0.1216    9 stem-height
 0.1149   13 stem-color
 0.0829    6 gill-attachment
 0.0829    7 gill-spacing
 0.0805   11 stem-root
 0.0748    4 cap-color
 0.0742    8 gill-color
 0.0637   15 veil-color
 0.0617   18 spore-print-color
 0.0614    2 cap-shape
 0.0576   16 has-ring
 0.0464   19 habitat
 0.046    20 season
 0.0446    3 cap-surface
 0.0325   17 ring-type
 0.0199    5 does-bruise-or-bleed
 0         14 veil-type

Selected attributes: 10,1,12,9,13,6,7,11,4,8,15,18,2,16,19,20,3,17,5,14 : 20
```

Cut-off value of 0.08 chosen - all attributes scoring below were discarded.

For our datasets with missing values as mean/mode:



```
10:56:16 - BestFirst + CfsSubsetEval
10:56:22 - Ranker + CorrelationAttributeEval
10:56:28 - Ranker + InfoGainAttributeEval
```

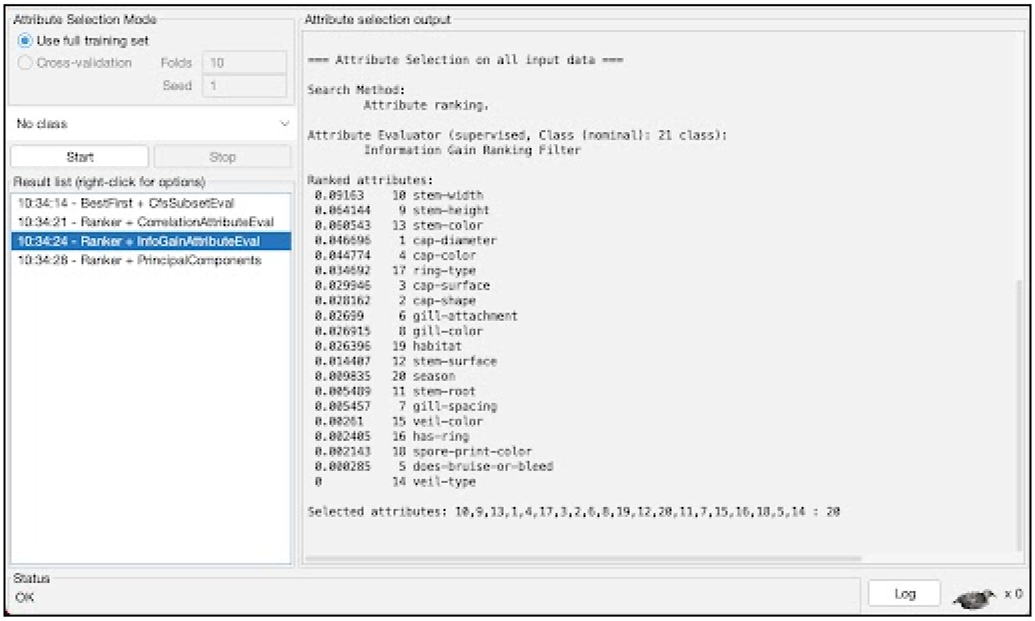
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Correlation Ranking Filter
Ranked attributes:
 0.1974   10 stem-width
 0.1811    1 cap-diameter
 0.1567   12 stem-surface
 0.1216    9 stem-height
 0.1149   13 stem-color
 0.0829    6 gill-attachment
 0.0829    7 gill-spacing
 0.0805   11 stem-root
 0.0748    4 cap-color
 0.0742    8 gill-color
 0.0637   15 veil-color
 0.0617   18 spore-print-color
 0.0614    2 cap-shape
 0.0576   16 has-ring
 0.0464   19 habitat
 0.046    20 season
 0.0446    3 cap-surface
 0.0325   17 ring-type
 0.0199    5 does-bruise-or-bleed
 0        14 veil-type

Selected attributes: 10,1,12,9,13,6,7,11,4,8,15,18,2,16,19,20,3,17,5,14 : 20
```

Cut-off value of 0.08 chosen.

c) InfoGainAttributeEval

Sorts attributes based on information gain/entropy loss.

For our datasets with missing values as "?":

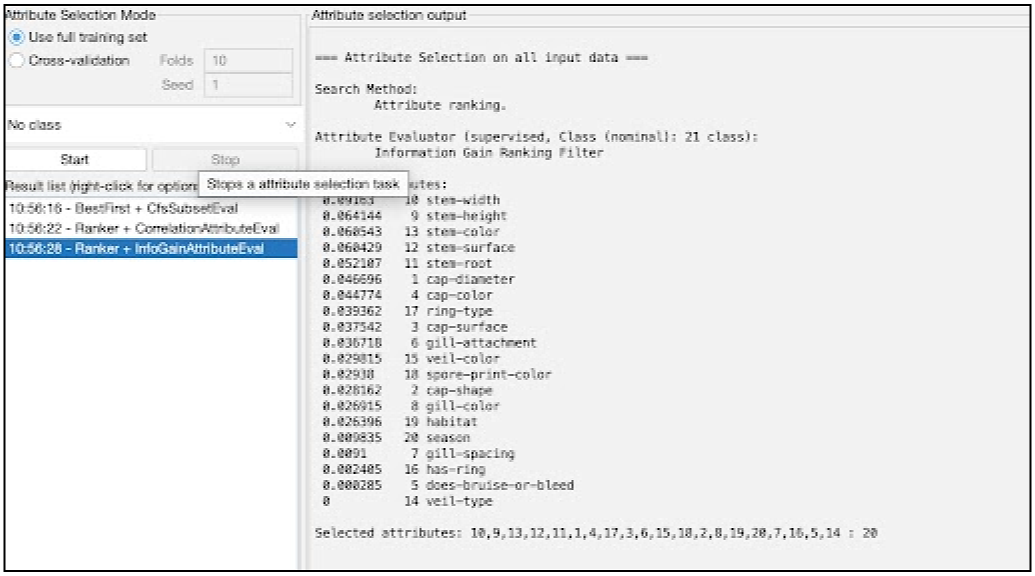


```
10:34:14 - BestFirst + CfsSubsetEval
10:34:21 - Ranker + CorrelationAttributeEval
10:34:24 - Ranker + InfoGainAttributeEval
10:34:28 - Ranker + PrincipalComponents
```

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.09163    10 stem-width
 0.064144    9 stem-height
 0.060543   13 stem-color
 0.046696    1 cap-diameter
 0.044774    4 cap-color
 0.034692   17 ring-type
 0.029946    3 cap-surface
 0.028162    2 cap-shape
 0.02699     6 gill-attachment
 0.026915    8 gill-color
 0.026396   19 habitat
 0.014407   12 stem-surface
 0.009835   20 season
 0.005409   11 stem-root
 0.005457    7 gill-spacing
 0.00261    15 veil-color
 0.002405   16 has-ring
 0.002143   18 spore-print-color
 0.000285    5 does-bruise-or-bleed
 0          14 veil-type

Selected attributes: 10,9,13,1,4,17,3,2,6,8,19,12,20,11,7,15,16,18,5,14 : 20
```

Cut-off value of 0.025 chosen.

For our datasets with missing values as mean/mode:


Attribute Selection Mode
● Use full training set
○ Cross-validation
Folds 10
Seed 1
No class
Start
Stop
Result list (right-click for options)
Stops a attribute selection task
10:56:16 - BestFirst + CfsSubsetEval
10:56:22 - Ranker + CorrelationAttributeEval
10:56:28 - Ranker + InfoGainAttributeEval
Attribute selection output

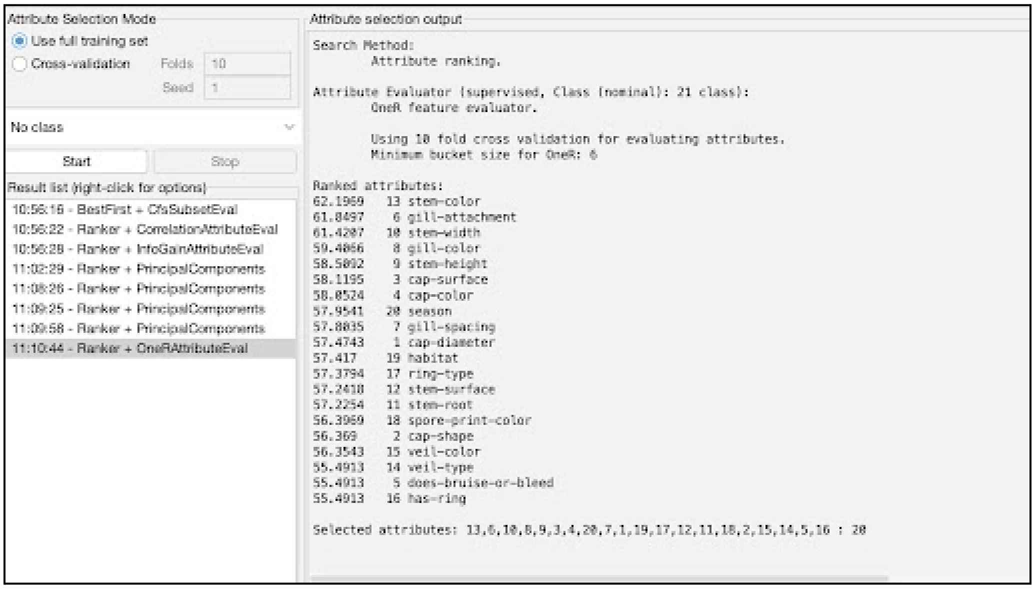
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        Information Gain Ranking Filter

utes:
 0.09163    10 stem-width
 0.064144    9 stem-height
 0.060543   13 stem-color
 0.060429   12 stem-surface
 0.052107   11 stem-root
 0.046696    1 cap-diameter
 0.044774    4 cap-color
 0.039362   17 ring-type
 0.037542    3 cap-surface
 0.036718    6 gill-attachment
 0.029815   15 veil-color
 0.02938    18 spore-print-color
 0.028162    2 cap-shape
 0.026915    8 gill-color
 0.026396   19 habitat
 0.009835   20 season
 0.0091      7 gill-spacing
 0.002405   16 has-ring
 0.000285    5 does-bruise-or-bleed
 0          14 veil-type

Selected attributes: 10,9,13,12,11,1,4,17,3,6,15,18,2,8,19,20,7,16,5,14 : 20
```

Cut-off value of 0.04 chosen.

d) OneRAttributeEval

Chooses attributes with OneR Classifier.

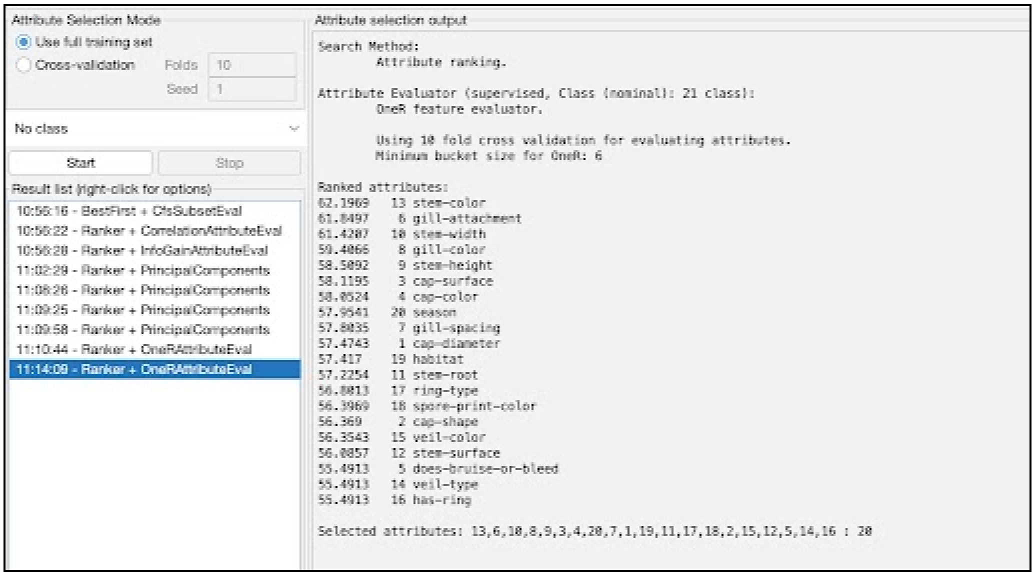
For our datasets with missing values as "?":


Attribute Selection Mode
● Use full training set
○ Cross-validation
Folds 10
Seed 1
No class
Start
Stop
Result list (right-click for options)
10:56:16 - BestFirst + CfsSubsetEval
10:56:22 - Ranker + CorrelationAttributeEval
10:56:28 - Ranker + InfoGainAttributeEval
11:02:29 - Ranker + PrincipalComponents
11:08:26 - Ranker + PrincipalComponents
11:09:25 - Ranker + PrincipalComponents
11:09:58 - Ranker + PrincipalComponents
11:10:44 - Ranker + OneRAttributeEval
Attribute selection output

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
62.1969   13 stem-color
61.8497    6 gill-attachment
61.4207   10 stem-width
59.4066    8 gill-color
58.5092    9 stem-height
58.1195    3 cap-surface
58.0524    4 cap-color
57.9541   20 season
57.8835    7 gill-spacing
57.4743    1 cap-diameter
57.417    19 habitat
57.3794   17 ring-type
57.2418   12 stem-surface
57.2254   11 stem-root
56.3969   18 spore-print-color
56.369     2 cap-shape
56.3543   15 veil-color
55.4913   14 veil-type
55.4913    5 does-bruise-or-bleed
55.4913   16 has-ring

Selected attributes: 13,6,10,8,9,3,4,20,7,1,19,17,12,11,18,2,15,14,5,16 : 20
```

Cut-off of 57.5 chosen.

For our datasets with missing values as mean/mode:



Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

No class

Result list (right-click for options)

```
10:56:16 - BestFirst + CfsSubsetEval
10:56:22 - Ranker + CorrelationAttributeEval
10:56:28 - Ranker + InfoGainAttributeEval
11:02:29 - Ranker + PrincipalComponents
11:08:26 - Ranker + PrincipalComponents
11:09:25 - Ranker + PrincipalComponents
11:09:58 - Ranker + PrincipalComponents
11:10:44 - Ranker + OneRAttributeEval
11:14:09 - Ranker + OneRAttributeEval
```

Attribute selection output

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
62.1969   13 stem-color
61.8497    6 gill-attachment
61.4207   10 stem-width
59.4066    8 gill-color
58.5092    9 stem-height
58.1195    3 cap-surface
58.0524    4 cap-color
57.9541   20 season
57.8035    7 gill-spacing
57.4743    1 cap-diameter
57.417    19 habitat
57.2254   11 stem-root
56.8013   17 ring-type
56.3969   18 spore-print-color
56.369     2 cap-shape
56.3543   15 veil-color
56.0857   12 stem-surface
55.4913    5 does-bruise-or-bleed
55.4913   14 veil-type
55.4913   16 has-ring

Selected attributes: 13,6,10,8,9,3,4,20,7,1,19,11,17,18,2,15,12,5,14,16 : 20
```

Cut-off value of 57.5 chosen.

e) Intuitive Selection

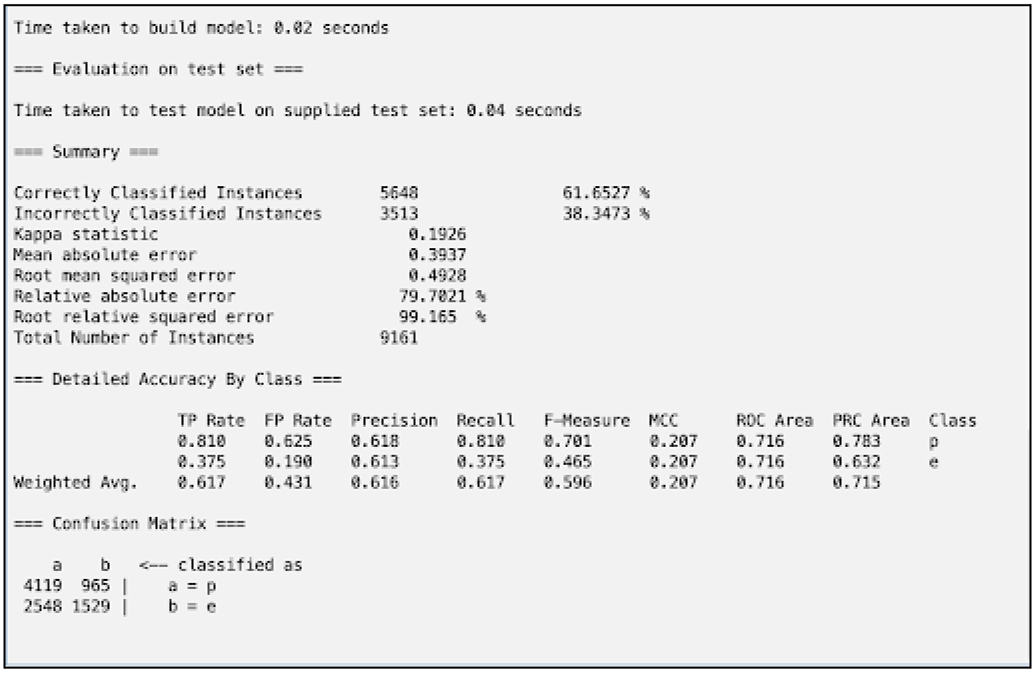
For both datasets, we selected 11 attributes:

1. Cap-shape
2. Cap-surface
3. Cap-color
4. Gill-color
5. Stem-surface
6. Stem-color
7. Veil-type
8. Veil-color
9. Has-ring
10. Ring-type
11. Spore-print-color

We chose these because poisonous plants and animals are usually identifiable by their shape and color, not necessarily their size. We thus removed the attributes that had to do with the measurements of the mushrooms, as well as some others like season and habitat.

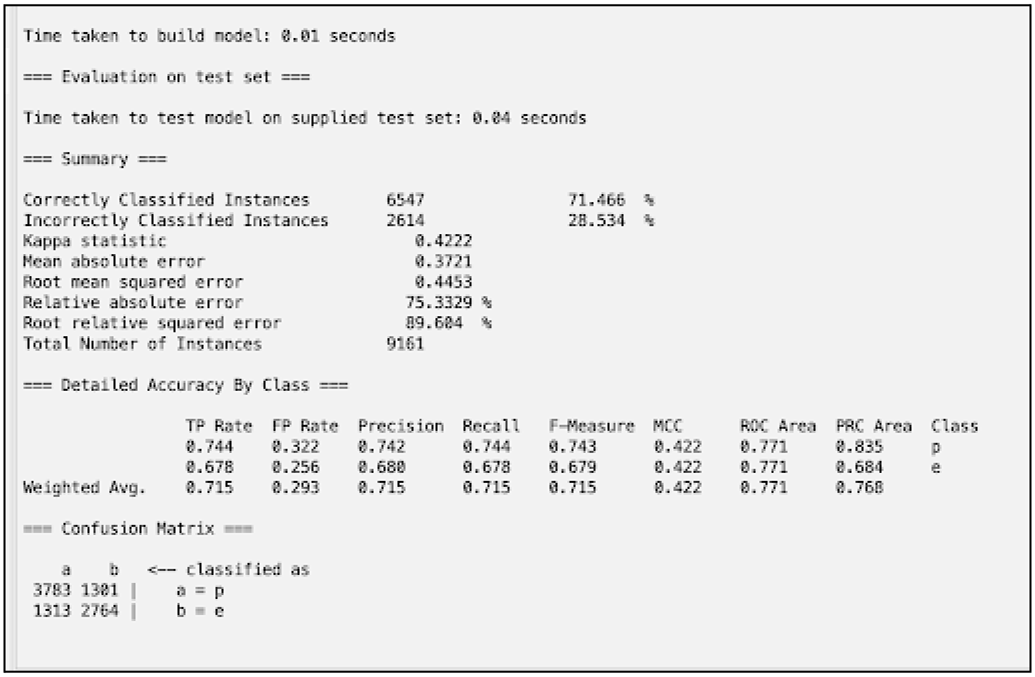
Data Mining Results & Evaluation

cfs_test.arff, naivebayes



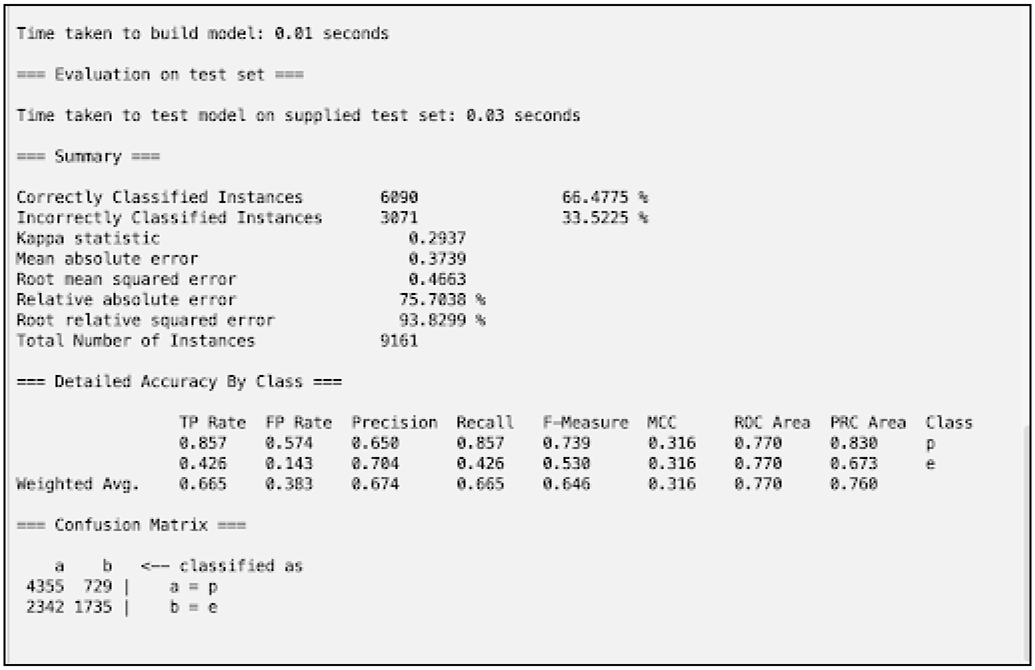
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        5648               61.6527 %
Incorrectly Classified Instances      3513               38.3473 %
Kappa statistic                          0.1926
Mean absolute error                      0.3937
Root mean squared error                  0.4928
Relative absolute error                 79.7021 %
Root relative squared error             99.165  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.810    0.625    0.618      0.810    0.701      0.207    0.716     0.783     p
                 0.375    0.190    0.613      0.375    0.465      0.207    0.716     0.632     e
Weighted Avg.    0.617    0.431    0.616      0.617    0.596      0.207    0.716     0.715

=== Confusion Matrix ===

    a    b   <-- classified as
 4119  965 |    a = p
 2548 1529 |    b = e
```

cfssubset_no_miss_test_set.arff, naivebayes



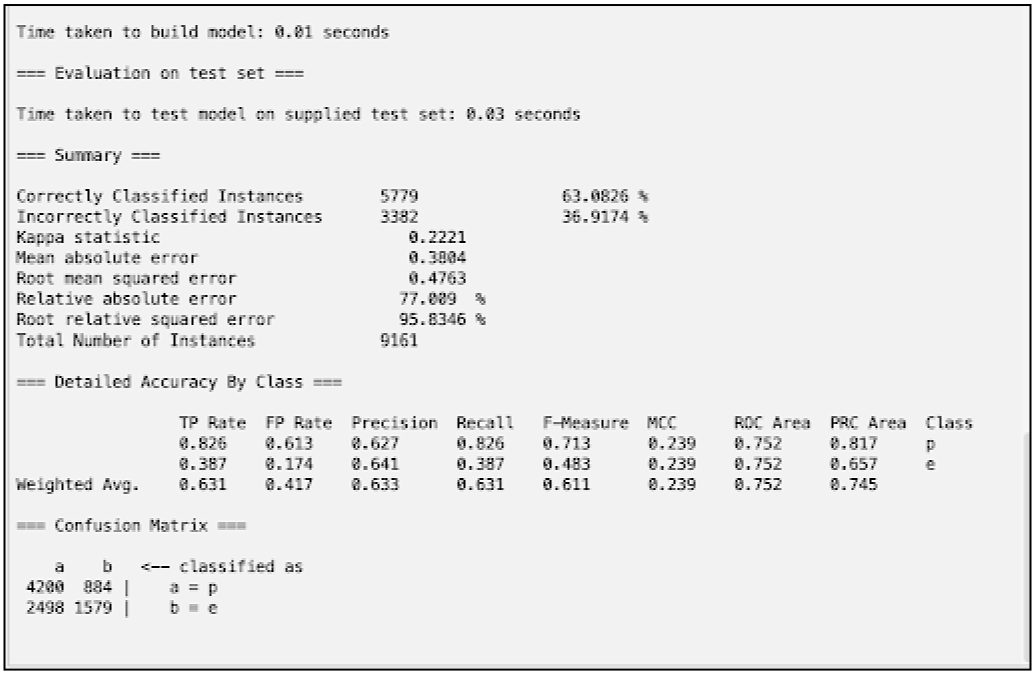
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        6547               71.466  %
Incorrectly Classified Instances      2614               28.534  %
Kappa statistic                          0.4222
Mean absolute error                      0.3721
Root mean squared error                  0.4453
Relative absolute error                 75.3329 %
Root relative squared error             89.604  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.744    0.322    0.742      0.744    0.743      0.422    0.771     0.835     p
                 0.678    0.256    0.680      0.678    0.679      0.422    0.771     0.684     e
Weighted Avg.    0.715    0.293    0.715      0.715    0.715      0.422    0.771     0.768

=== Confusion Matrix ===

    a    b   <-- classified as
 3783 1301 |    a = p
 1313 2764 |    b = e
```

correlation_no_miss_test_set.arff, naivebayes



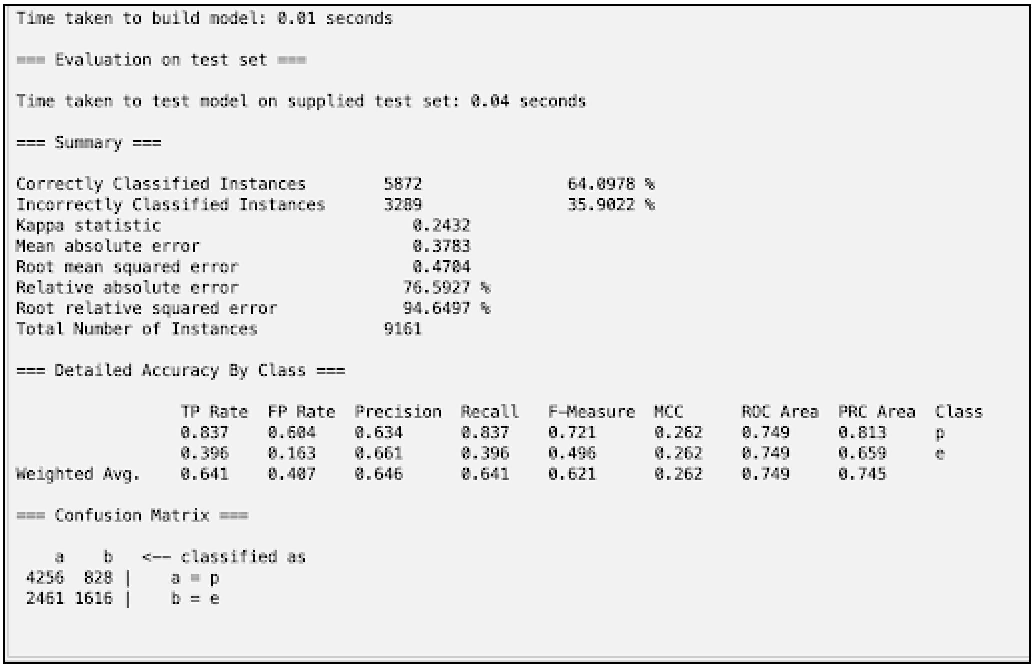
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        6090               66.4775 %
Incorrectly Classified Instances      3071               33.5225 %
Kappa statistic                          0.2937
Mean absolute error                      0.3739
Root mean squared error                  0.4663
Relative absolute error                 75.7038 %
Root relative squared error             93.8299 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.857    0.574    0.650      0.857    0.739      0.316    0.770     0.830     p
                 0.426    0.143    0.704      0.426    0.530      0.316    0.770     0.673     e
Weighted Avg.    0.665    0.383    0.674      0.665    0.646      0.316    0.770     0.760

=== Confusion Matrix ===

    a    b   <-- classified as
 4355  729 |    a = p
 2342 1735 |    b = e
```

correlation_test_set.arff, naivebayes



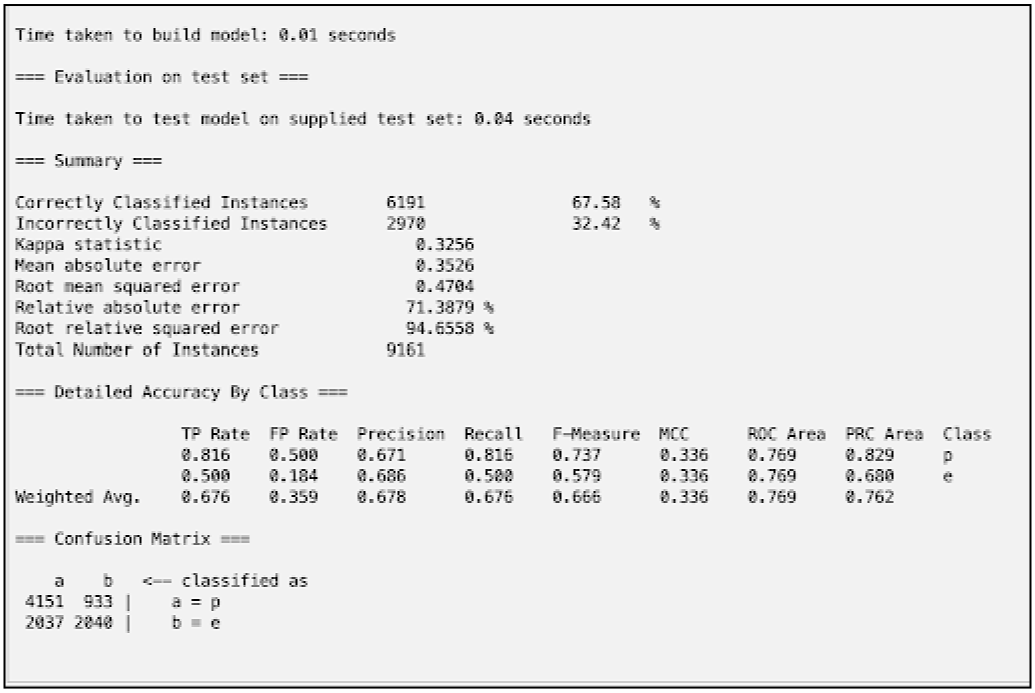
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        5779               63.0826 %
Incorrectly Classified Instances      3382               36.9174 %
Kappa statistic                          0.2221
Mean absolute error                      0.3804
Root mean squared error                  0.4763
Relative absolute error                 77.009  %
Root relative squared error             95.8346 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.826    0.613    0.627      0.826    0.713      0.239    0.752     0.817     p
                 0.387    0.174    0.641      0.387    0.483      0.239    0.752     0.657     e
Weighted Avg.    0.631    0.417    0.633      0.631    0.611      0.239    0.752     0.745

=== Confusion Matrix ===

    a    b   <-- classified as
 4200  884 |    a = p
 2498 1579 |    b = e
```

infogain_no_miss.arff, naivebayes



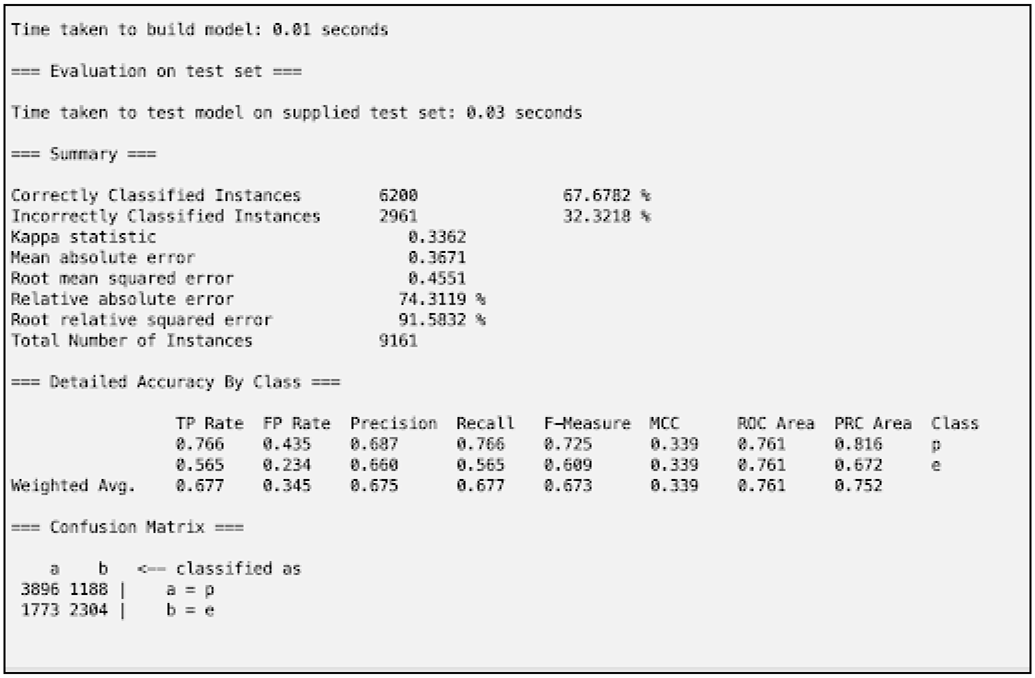
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        5872               64.0978 %
Incorrectly Classified Instances      3289               35.9022 %
Kappa statistic                          0.2432
Mean absolute error                      0.3783
Root mean squared error                  0.4704
Relative absolute error                 76.5927 %
Root relative squared error             94.6497 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.837    0.604    0.634      0.837    0.721      0.262    0.749     0.813     p
                 0.396    0.163    0.661      0.396    0.496      0.262    0.749     0.659     e
Weighted Avg.    0.641    0.407    0.646      0.641    0.621      0.262    0.749     0.745

=== Confusion Matrix ===

    a    b   <-- classified as
 4256  828 |    a = p
 2461 1616 |    b = e
```

infogain_test_set.arff, naivebayes



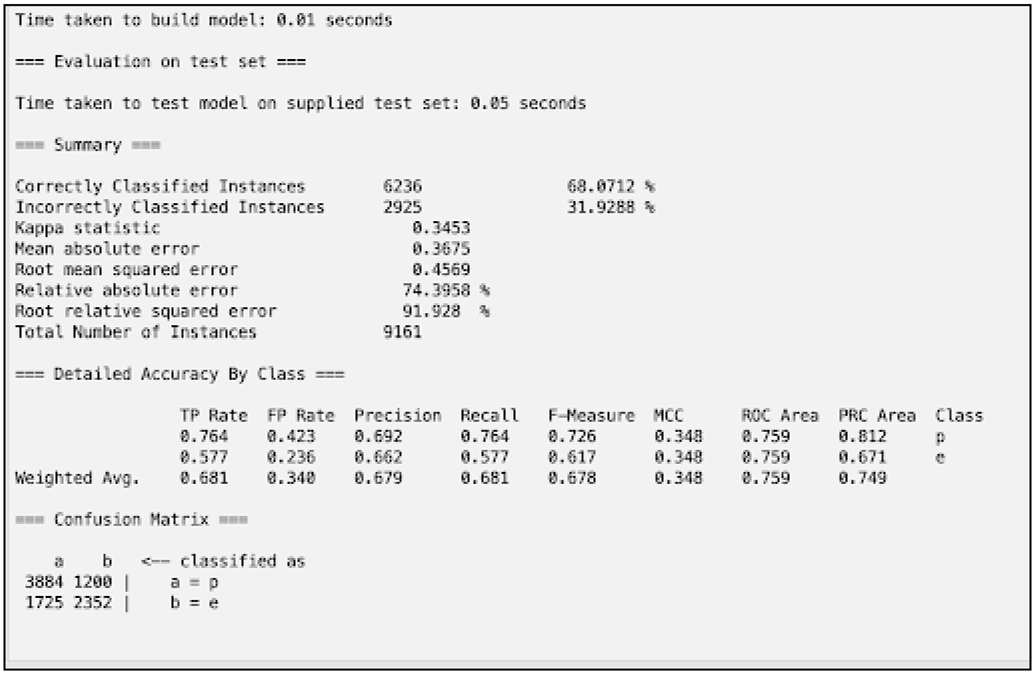
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.04 seconds

=== Summary ===

Correctly Classified Instances        6191               67.58   %
Incorrectly Classified Instances      2970               32.42   %
Kappa statistic                          0.3256
Mean absolute error                      0.3526
Root mean squared error                  0.4704
Relative absolute error                 71.3879 %
Root relative squared error             94.6558 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.816    0.500    0.671      0.816    0.737      0.336    0.769     0.829     p
                 0.500    0.184    0.686      0.500    0.579      0.336    0.769     0.680     e
Weighted Avg.    0.676    0.359    0.678      0.676    0.666      0.336    0.769     0.762

=== Confusion Matrix ===

    a    b   <-- classified as
 4151  933 |    a = p
 2037 2040 |    b = e
```

oner_no_miss_test_set.arff, naivebayes



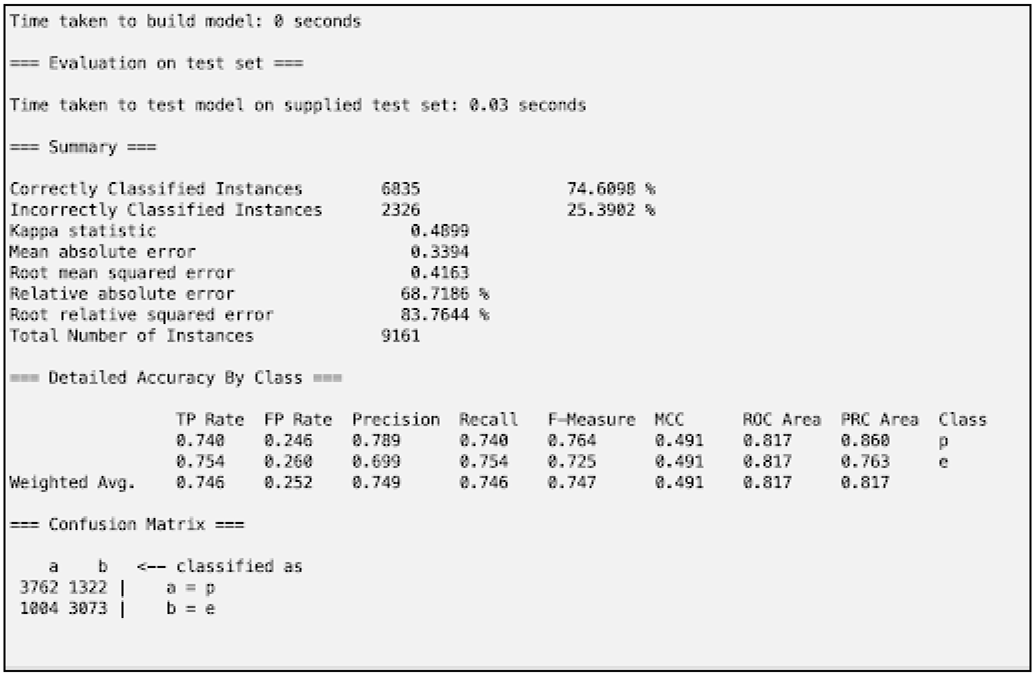
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        6200               67.6782 %
Incorrectly Classified Instances      2961               32.3218 %
Kappa statistic                          0.3362
Mean absolute error                      0.3671
Root mean squared error                  0.4551
Relative absolute error                 74.3119 %
Root relative squared error             91.5832 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.766    0.435    0.687      0.766    0.725      0.339    0.761     0.816     p
                 0.565    0.234    0.660      0.565    0.609      0.339    0.761     0.672     e
Weighted Avg.    0.677    0.345    0.675      0.677    0.673      0.339    0.761     0.752

=== Confusion Matrix ===

    a    b   <-- classified as
 3896 1188 |    a = p
 1773 2304 |    b = e
```

oner_test_set.arff, naivebayes



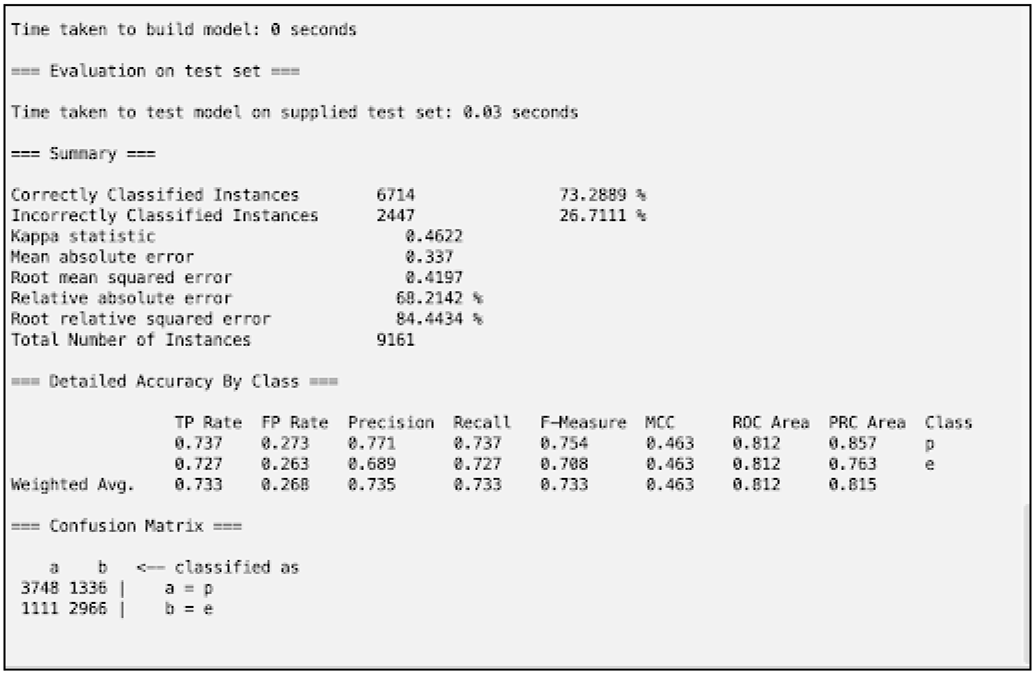
```
Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.05 seconds

=== Summary ===

Correctly Classified Instances        6236               68.0712 %
Incorrectly Classified Instances      2925               31.9288 %
Kappa statistic                          0.3453
Mean absolute error                      0.3675
Root mean squared error                  0.4569
Relative absolute error                 74.3958 %
Root relative squared error             91.928  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.764    0.423    0.692      0.764    0.726      0.348    0.759     0.812     p
                 0.577    0.236    0.662      0.577    0.617      0.348    0.759     0.671     e
Weighted Avg.    0.681    0.340    0.679      0.681    0.678      0.348    0.759     0.749

=== Confusion Matrix ===

    a    b   <-- classified as
 3884 1200 |    a = p
 1725 2352 |    b = e
```

selected_no_miss_test_set.arff, naivebayes



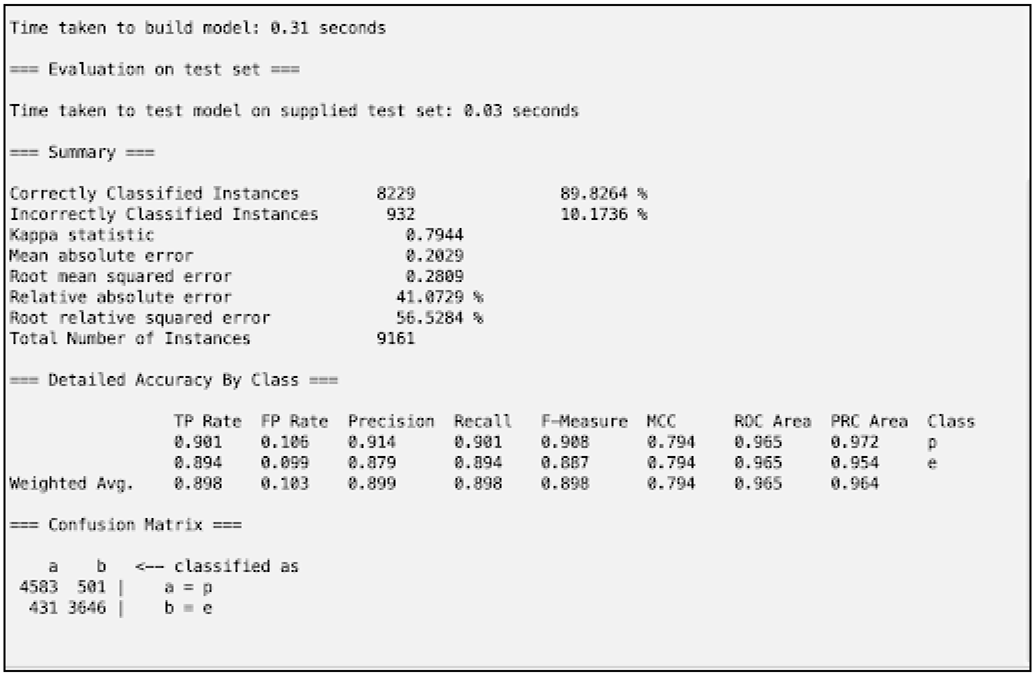
```
Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        6835               74.6098 %
Incorrectly Classified Instances      2326               25.3902 %
Kappa statistic                          0.4899
Mean absolute error                      0.3394
Root mean squared error                  0.4163
Relative absolute error                 68.7186 %
Root relative squared error             83.7644 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.740    0.246    0.789      0.740    0.764      0.491    0.817     0.860     p
                 0.754    0.260    0.699      0.754    0.725      0.491    0.817     0.763     e
Weighted Avg.    0.746    0.252    0.749      0.746    0.747      0.491    0.817     0.817

=== Confusion Matrix ===

    a    b   <-- classified as
 3762 1322 |    a = p
 1004 3073 |    b = e
```

selected_test_set.arff, naivebayes



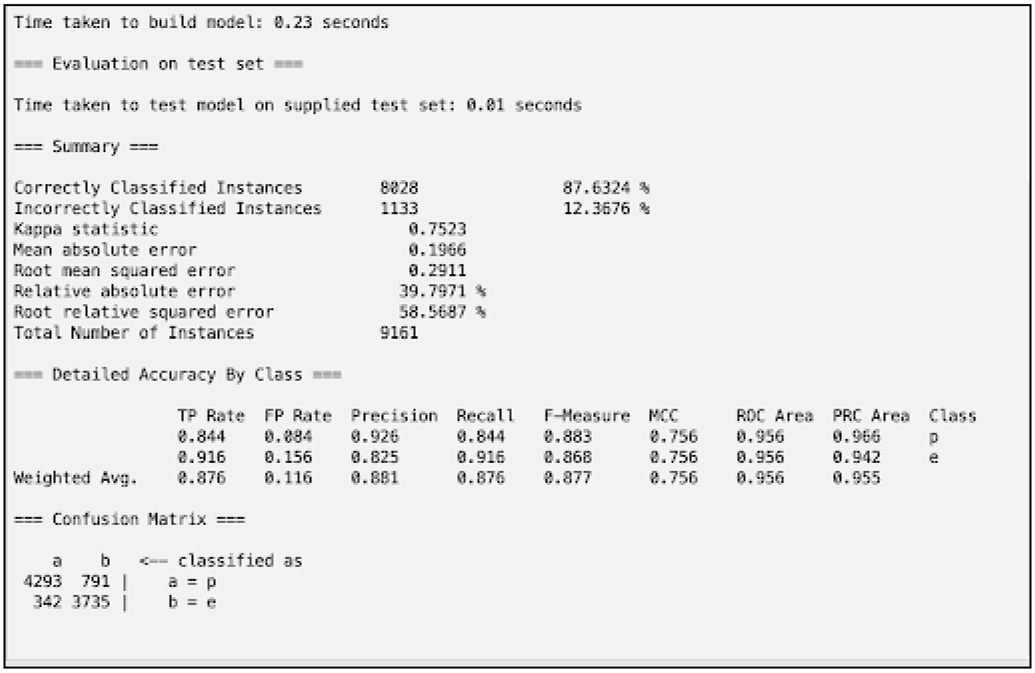
```
Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        6714               73.2889 %
Incorrectly Classified Instances      2447               26.7111 %
Kappa statistic                          0.4622
Mean absolute error                      0.337
Root mean squared error                  0.4197
Relative absolute error                 68.2142 %
Root relative squared error             84.4434 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.737    0.273    0.771      0.737    0.754      0.463    0.812     0.857     p
                 0.727    0.263    0.689      0.727    0.708      0.463    0.812     0.763     e
Weighted Avg.    0.733    0.268    0.735      0.733    0.733      0.463    0.812     0.815

=== Confusion Matrix ===

    a    b   <-- classified as
 3748 1336 |    a = p
 1111 2966 |    b = e
```

cfs_test_set.arff, DecisionTable



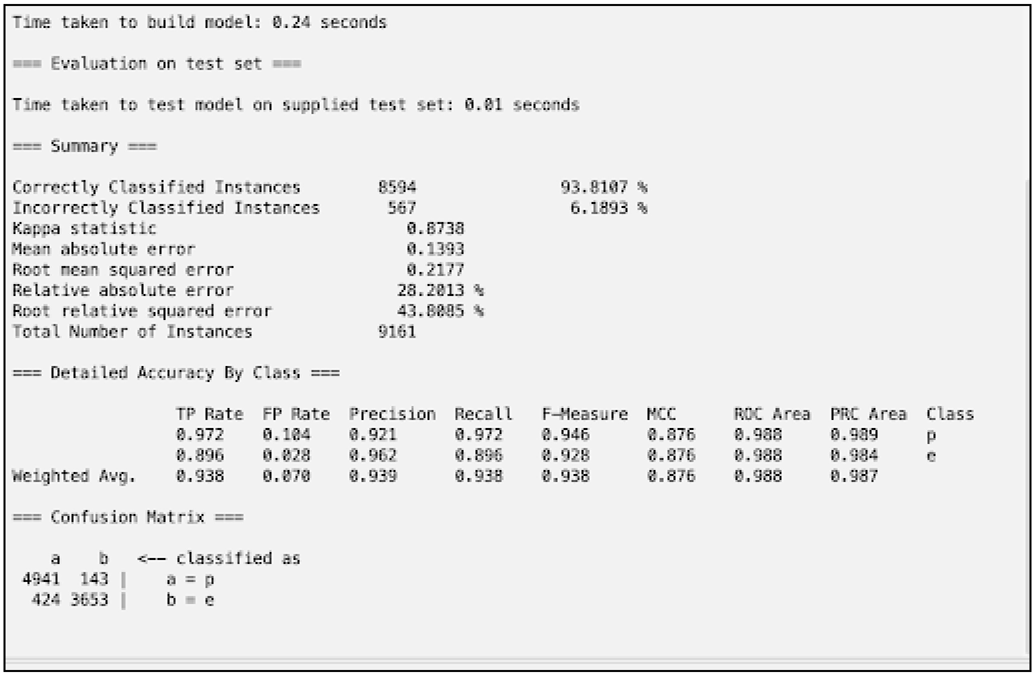
```
Time taken to build model: 0.31 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        8229               89.8264 %
Incorrectly Classified Instances       932               10.1736 %
Kappa statistic                          0.7944
Mean absolute error                      0.2029
Root mean squared error                  0.2809
Relative absolute error                 41.0729 %
Root relative squared error             56.5284 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.901    0.106    0.914      0.901    0.908      0.794    0.965     0.972     p
                 0.894    0.099    0.879      0.894    0.887      0.794    0.965     0.954     e
Weighted Avg.    0.898    0.103    0.899      0.898    0.898      0.794    0.965     0.964

=== Confusion Matrix ===

    a    b   <-- classified as
 4583  501 |    a = p
  431 3646 |    b = e
```

cfssubset_no_miss_test_set.arff, DecisionTable



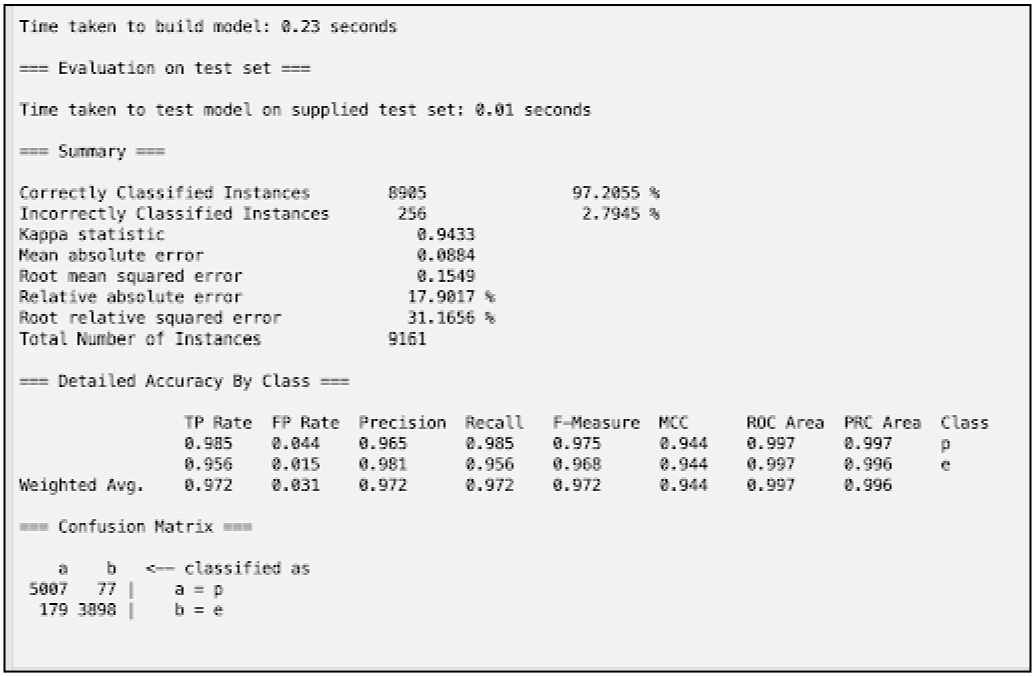
```
Time taken to build model: 0.23 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8028               87.6324 %
Incorrectly Classified Instances      1133               12.3676 %
Kappa statistic                          0.7523
Mean absolute error                      0.1966
Root mean squared error                  0.2911
Relative absolute error                 39.7971 %
Root relative squared error             58.5687 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.844    0.084    0.926      0.844    0.883      0.756    0.956     0.966     p
                 0.916    0.156    0.825      0.916    0.868      0.756    0.956     0.942     e
Weighted Avg.    0.876    0.116    0.881      0.876    0.877      0.756    0.956     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
 4293  791 |    a = p
  342 3735 |    b = e
```

correlation_no_miss_test_set.arff, DecisionTable



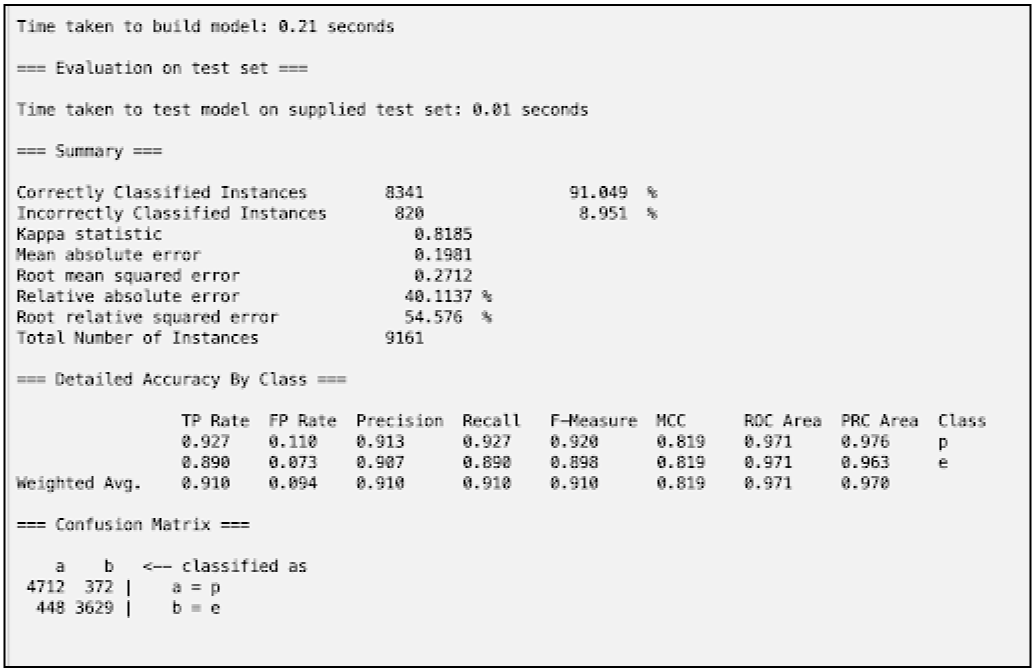
```
Time taken to build model: 0.24 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8594               93.8107 %
Incorrectly Classified Instances       567                6.1893 %
Kappa statistic                          0.8738
Mean absolute error                      0.1393
Root mean squared error                  0.2177
Relative absolute error                 28.2013 %
Root relative squared error             43.8085 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.972    0.104    0.921      0.972    0.946      0.876    0.988     0.989     p
                 0.896    0.028    0.962      0.896    0.928      0.876    0.988     0.984     e
Weighted Avg.    0.938    0.070    0.939      0.938    0.938      0.876    0.988     0.987

=== Confusion Matrix ===

    a    b   <-- classified as
 4941  143 |    a = p
  424 3653 |    b = e
```

correlation_test_set.arff, DecisionTable



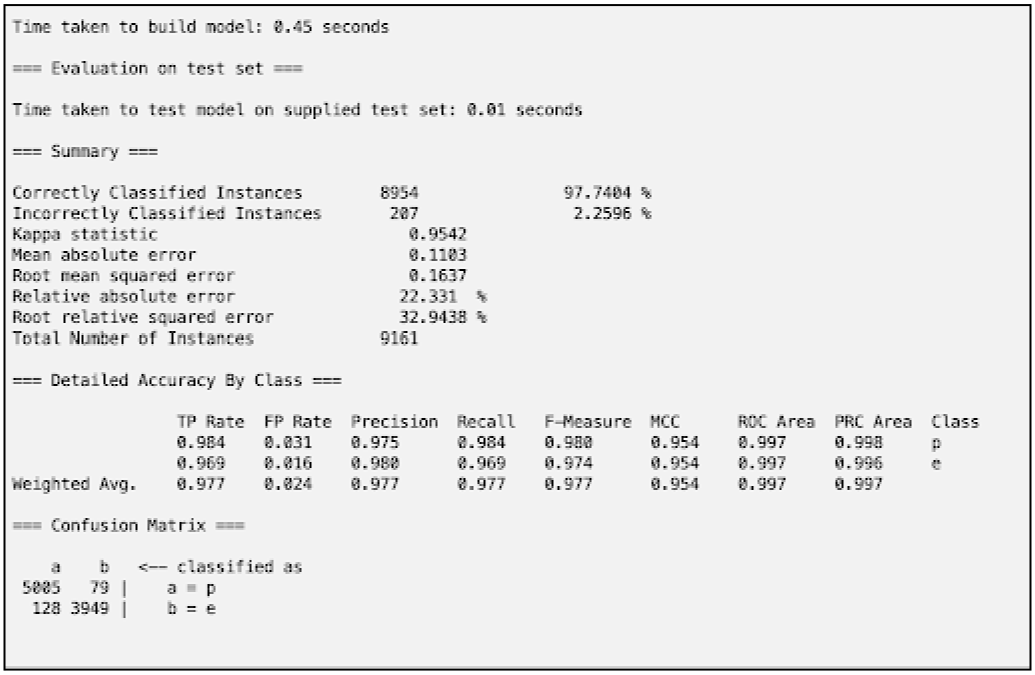
```
Time taken to build model: 0.23 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8905               97.2055 %
Incorrectly Classified Instances       256                2.7945 %
Kappa statistic                          0.9433
Mean absolute error                      0.0884
Root mean squared error                  0.1549
Relative absolute error                 17.9017 %
Root relative squared error             31.1656 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.985    0.044    0.965      0.985    0.975      0.944    0.997     0.997     p
                 0.956    0.015    0.981      0.956    0.968      0.944    0.997     0.996     e
Weighted Avg.    0.972    0.031    0.972      0.972    0.972      0.944    0.997     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 5007   77 |    a = p
  179 3898 |    b = e
```

infogain_no_miss_test_set.arff, DecisionTable



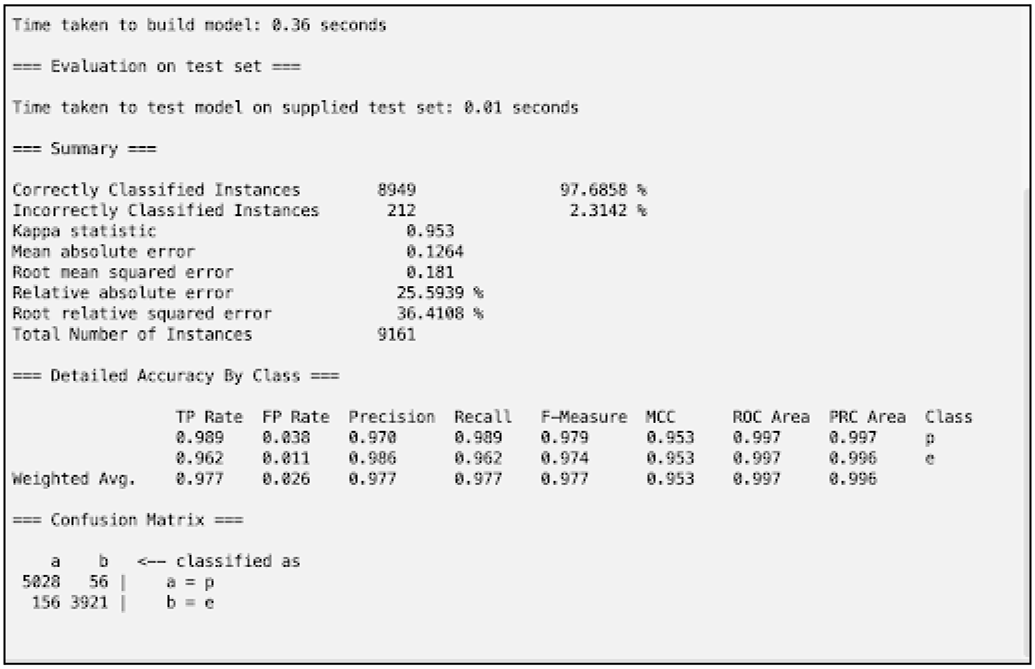
```
Time taken to build model: 0.21 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8341               91.049  %
Incorrectly Classified Instances       820                8.951  %
Kappa statistic                          0.8185
Mean absolute error                      0.1981
Root mean squared error                  0.2712
Relative absolute error                 40.1137 %
Root relative squared error             54.576  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.927    0.110    0.913      0.927    0.920      0.819    0.971     0.976     p
                 0.890    0.073    0.907      0.890    0.898      0.819    0.971     0.963     e
Weighted Avg.    0.910    0.094    0.910      0.910    0.910      0.819    0.971     0.970

=== Confusion Matrix ===

    a    b   <-- classified as
 4712  372 |    a = p
  448 3629 |    b = e
```

infogain_test_set.arff, DecisionTable



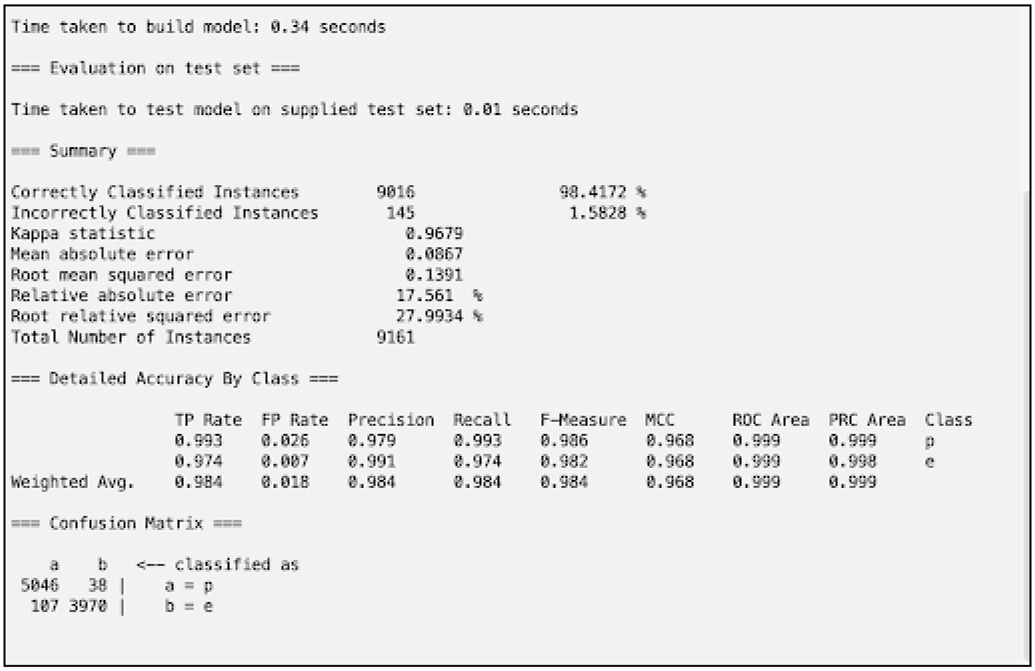
```
Time taken to build model: 0.45 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8954               97.7404 %
Incorrectly Classified Instances       207                2.2596 %
Kappa statistic                          0.9542
Mean absolute error                      0.1103
Root mean squared error                  0.1637
Relative absolute error                 22.331  %
Root relative squared error             32.9438 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.031    0.975      0.984    0.980      0.954    0.997     0.998     p
                 0.969    0.016    0.980      0.969    0.974      0.954    0.997     0.996     e
Weighted Avg.    0.977    0.024    0.977      0.977    0.977      0.954    0.997     0.997

=== Confusion Matrix ===

    a    b   <-- classified as
 5005   79 |    a = p
  128 3949 |    b = e
```

oner_no_miss_test_set.arff, DecisionTable



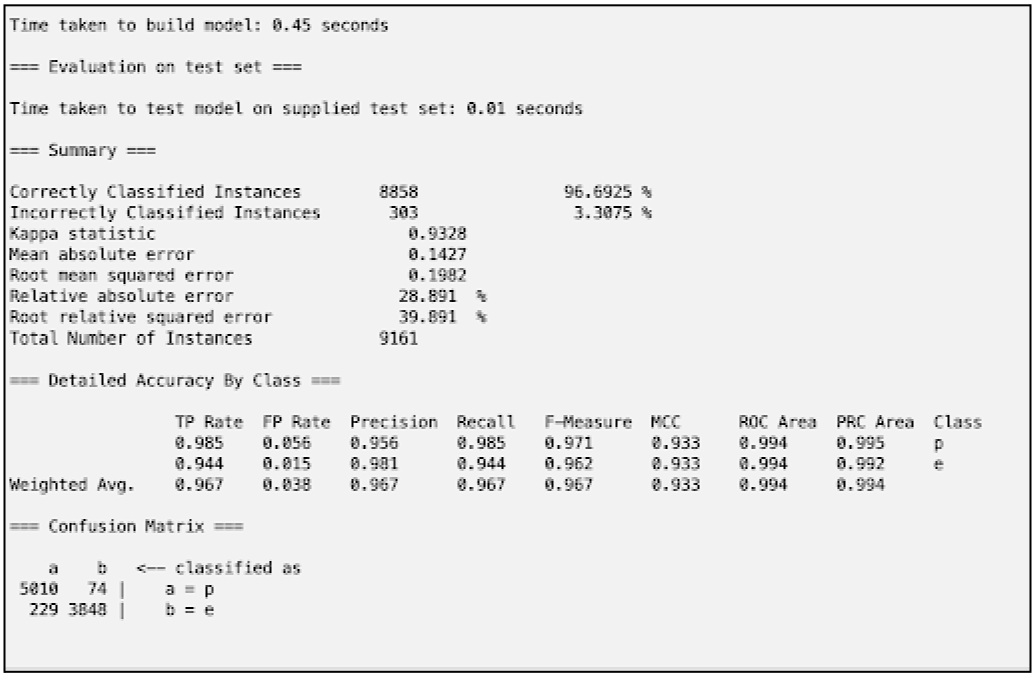
```
Time taken to build model: 0.36 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8949               97.6858 %
Incorrectly Classified Instances       212                2.3142 %
Kappa statistic                          0.953
Mean absolute error                      0.1264
Root mean squared error                  0.181
Relative absolute error                 25.5939 %
Root relative squared error             36.4108 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.989    0.038    0.970      0.989    0.979      0.953    0.997     0.997     p
                 0.962    0.011    0.986      0.962    0.974      0.953    0.997     0.996     e
Weighted Avg.    0.977    0.026    0.977      0.977    0.977      0.953    0.997     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 5028   56 |    a = p
  156 3921 |    b = e
```

oner_test_set.arff, DecisionTable



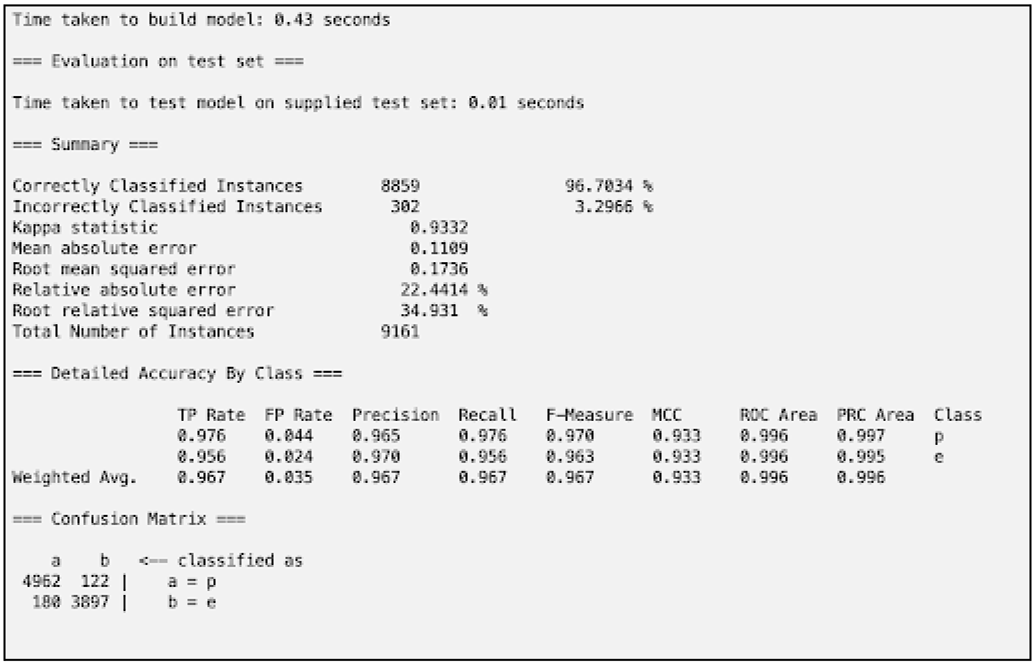
```
Time taken to build model: 0.34 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        9016               98.4172 %
Incorrectly Classified Instances       145                1.5828 %
Kappa statistic                          0.9679
Mean absolute error                      0.0867
Root mean squared error                  0.1391
Relative absolute error                 17.561  %
Root relative squared error             27.9934 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.993    0.026    0.979      0.993    0.986      0.968    0.999     0.999     p
                 0.974    0.007    0.991      0.974    0.982      0.968    0.999     0.998     e
Weighted Avg.    0.984    0.018    0.984      0.984    0.984      0.968    0.999     0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 5046   38 |    a = p
  107 3970 |    b = e
```

selected_no_miss_test_set.arff, DecisionTable



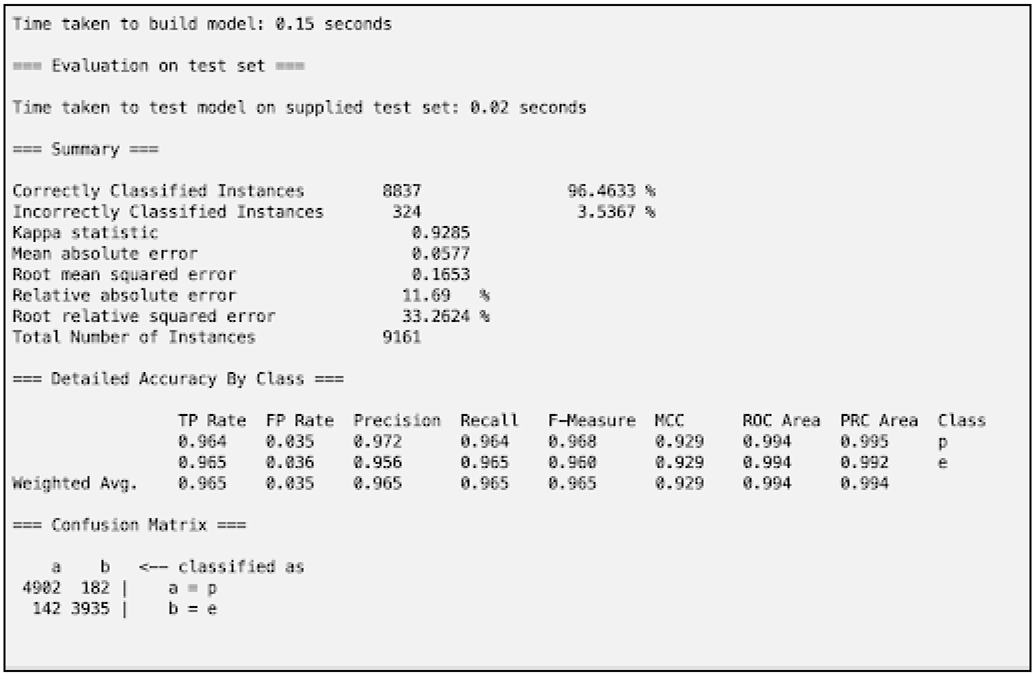
```
Time taken to build model: 0.45 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8858               96.6925 %
Incorrectly Classified Instances       303                3.3075 %
Kappa statistic                          0.9328
Mean absolute error                      0.1427
Root mean squared error                  0.1982
Relative absolute error                 28.891  %
Root relative squared error             39.891  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.985    0.056    0.956      0.985    0.971      0.933    0.994     0.995     p
                 0.944    0.015    0.981      0.944    0.962      0.933    0.994     0.992     e
Weighted Avg.    0.967    0.038    0.967      0.967    0.967      0.933    0.994     0.994

=== Confusion Matrix ===

    a    b   <-- classified as
 5010   74 |    a = p
  229 3848 |    b = e
```

selected_test_set.arff, DecisionTable



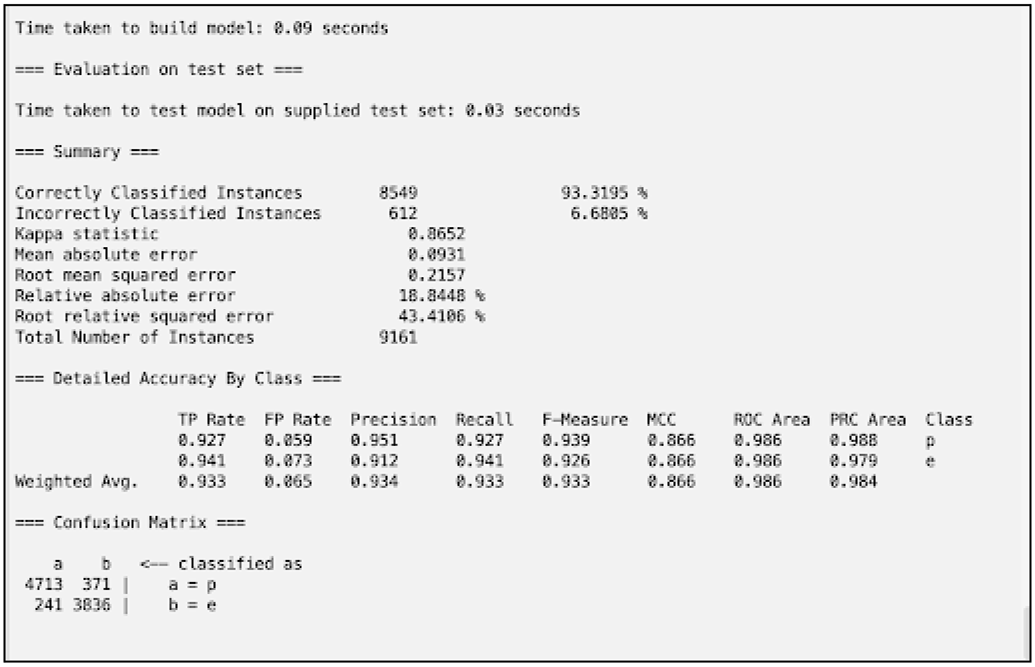
```
Time taken to build model: 0.43 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8859               96.7034 %
Incorrectly Classified Instances       302                3.2966 %
Kappa statistic                          0.9332
Mean absolute error                      0.1109
Root mean squared error                  0.1736
Relative absolute error                 22.4414 %
Root relative squared error             34.931  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976    0.044    0.965      0.976    0.970      0.933    0.996     0.997     p
                 0.956    0.024    0.970      0.956    0.963      0.933    0.996     0.995     e
Weighted Avg.    0.967    0.035    0.967      0.967    0.967      0.933    0.996     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 4962  122 |    a = p
  180 3897 |    b = e
```

cfs_test_set.arff, J48



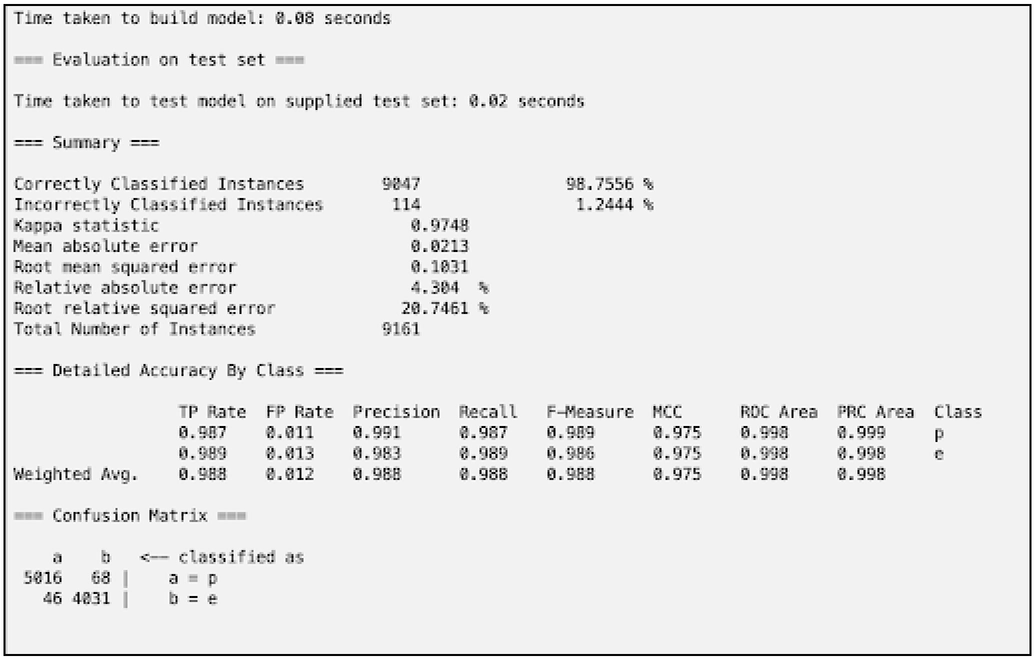
```
Time taken to build model: 0.15 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        8837               96.4633 %
Incorrectly Classified Instances       324                3.5367 %
Kappa statistic                          0.9285
Mean absolute error                      0.0577
Root mean squared error                  0.1653
Relative absolute error                 11.69   %
Root relative squared error             33.2624 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.035    0.972      0.964    0.968      0.929    0.994     0.995     p
                 0.965    0.036    0.956      0.965    0.960      0.929    0.994     0.992     e
Weighted Avg.    0.965    0.035    0.965      0.965    0.965      0.929    0.994     0.994

=== Confusion Matrix ===

    a    b   <-- classified as
 4902  182 |    a = p
  142 3935 |    b = e
```

cfssubset_no_miss_test_set.arff, J48



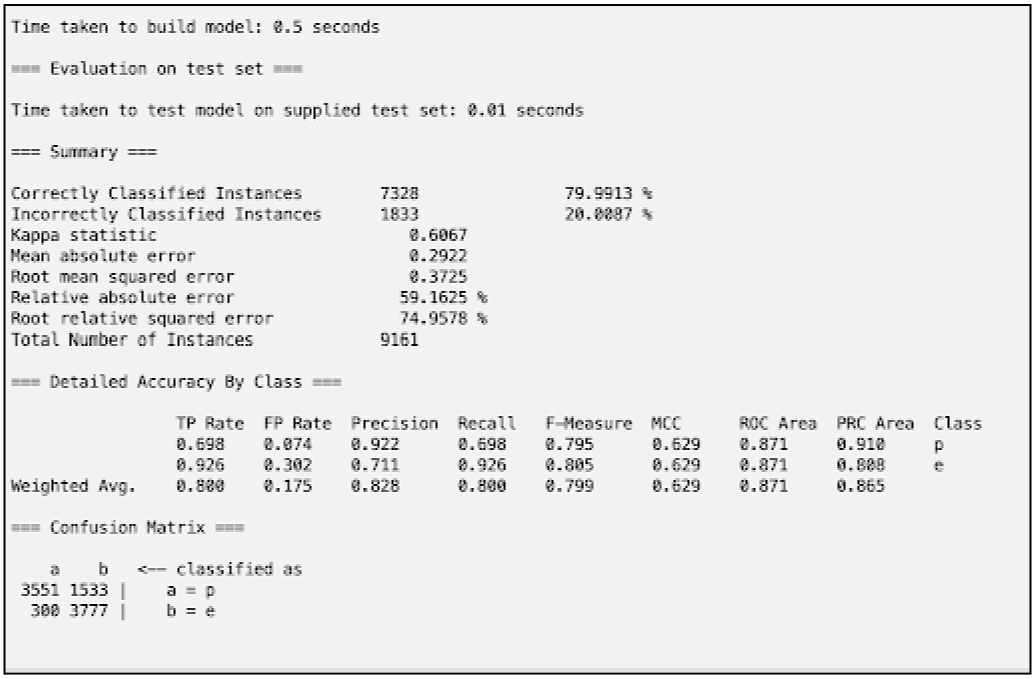
```
Time taken to build model: 0.09 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances        8549               93.3195 %
Incorrectly Classified Instances       612                6.6805 %
Kappa statistic                          0.8652
Mean absolute error                      0.0931
Root mean squared error                  0.2157
Relative absolute error                 18.8448 %
Root relative squared error             43.4106 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.927    0.059    0.951      0.927    0.939      0.866    0.986     0.988     p
                 0.941    0.073    0.912      0.941    0.926      0.866    0.986     0.979     e
Weighted Avg.    0.933    0.065    0.934      0.933    0.933      0.866    0.986     0.984

=== Confusion Matrix ===

    a    b   <-- classified as
 4713  371 |    a = p
  241 3836 |    b = e
```

correlation_no_miss_test_set.arff, J48



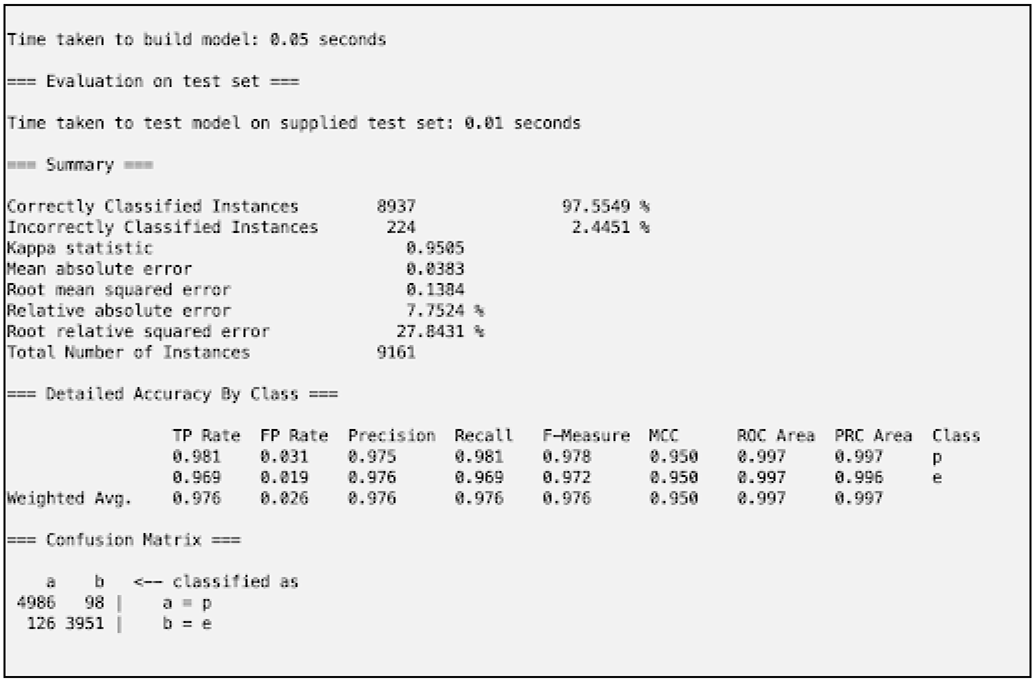
```
Time taken to build model: 0.08 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        9047               98.7556 %
Incorrectly Classified Instances       114                1.2444 %
Kappa statistic                          0.9748
Mean absolute error                      0.0213
Root mean squared error                  0.1031
Relative absolute error                  4.304  %
Root relative squared error             20.7461 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.987    0.011    0.991      0.987    0.989      0.975    0.998     0.999     p
                 0.989    0.013    0.983      0.989    0.986      0.975    0.998     0.998     e
Weighted Avg.    0.988    0.012    0.988      0.988    0.988      0.975    0.998     0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 5016   68 |    a = p
   46 4031 |    b = e
```

correlation_test_set.arff, J48



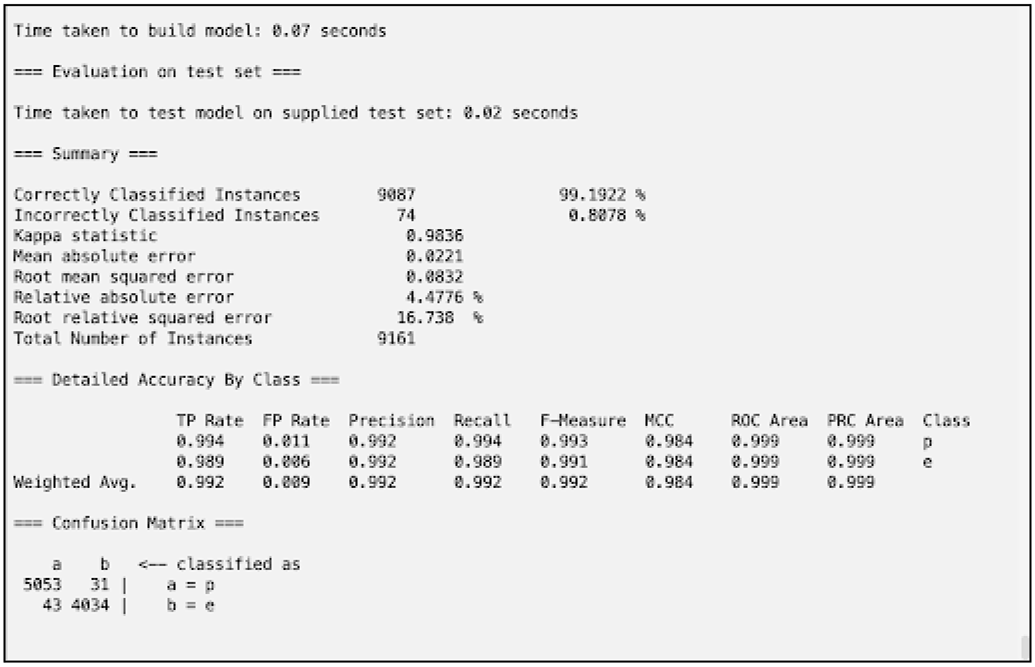
```
Time taken to build model: 0.5 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        7328               79.9913 %
Incorrectly Classified Instances      1833               20.0087 %
Kappa statistic                          0.6067
Mean absolute error                      0.2922
Root mean squared error                  0.3725
Relative absolute error                 59.1625 %
Root relative squared error             74.9578 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.698    0.074    0.922      0.698    0.795      0.629    0.871     0.910     p
                 0.926    0.302    0.711      0.926    0.805      0.629    0.871     0.808     e
Weighted Avg.    0.800    0.175    0.828      0.800    0.799      0.629    0.871     0.865

=== Confusion Matrix ===

    a    b   <-- classified as
 3551 1533 |    a = p
  300 3777 |    b = e
```

infogain_no_miss_test_set.arff, J48



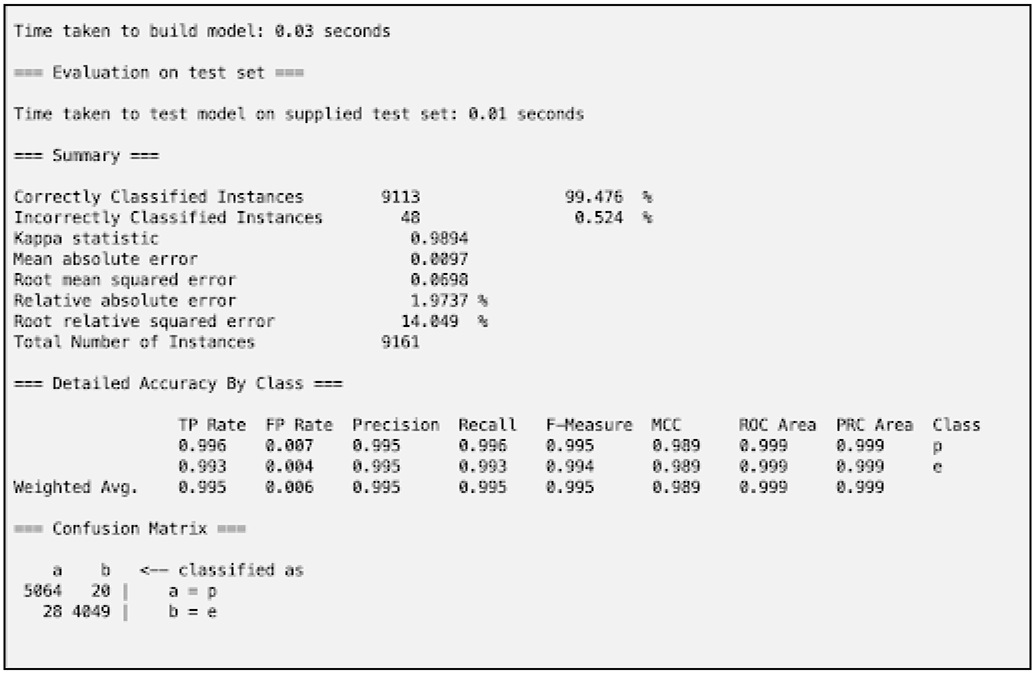
```
Time taken to build model: 0.05 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8937               97.5549 %
Incorrectly Classified Instances       224                2.4451 %
Kappa statistic                          0.9505
Mean absolute error                      0.0383
Root mean squared error                  0.1384
Relative absolute error                  7.7524 %
Root relative squared error             27.8431 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.031    0.975      0.981    0.978      0.950    0.997     0.997     p
                 0.969    0.019    0.976      0.969    0.972      0.950    0.997     0.996     e
Weighted Avg.    0.976    0.026    0.976      0.976    0.976      0.950    0.997     0.997

=== Confusion Matrix ===

    a    b   <-- classified as
 4986   98 |    a = p
  126 3951 |    b = e
```

infogain_test_set.arff, J48



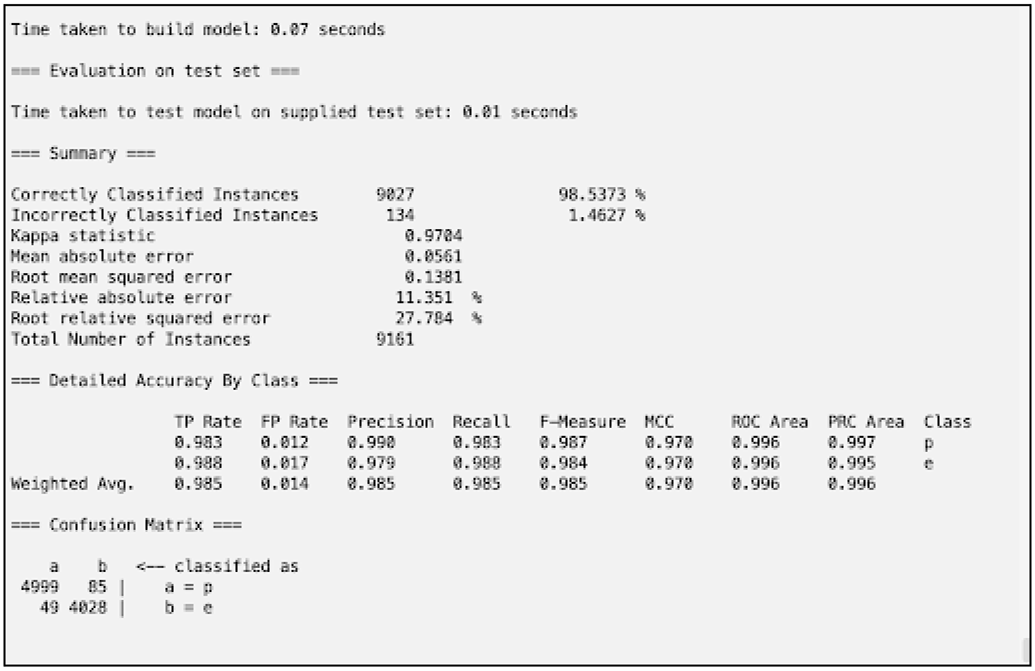
```
Time taken to build model: 0.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances        9087               99.1922 %
Incorrectly Classified Instances        74                0.8078 %
Kappa statistic                          0.9836
Mean absolute error                      0.0221
Root mean squared error                  0.0832
Relative absolute error                  4.4776 %
Root relative squared error             16.738  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.994    0.011    0.992      0.994    0.993      0.984    0.999     0.999     p
                 0.989    0.006    0.992      0.989    0.991      0.984    0.999     0.999     e
Weighted Avg.    0.992    0.009    0.992      0.992    0.992      0.984    0.999     0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 5053   31 |    a = p
   43 4034 |    b = e
```

oner_no_miss_test_set.arff, J48



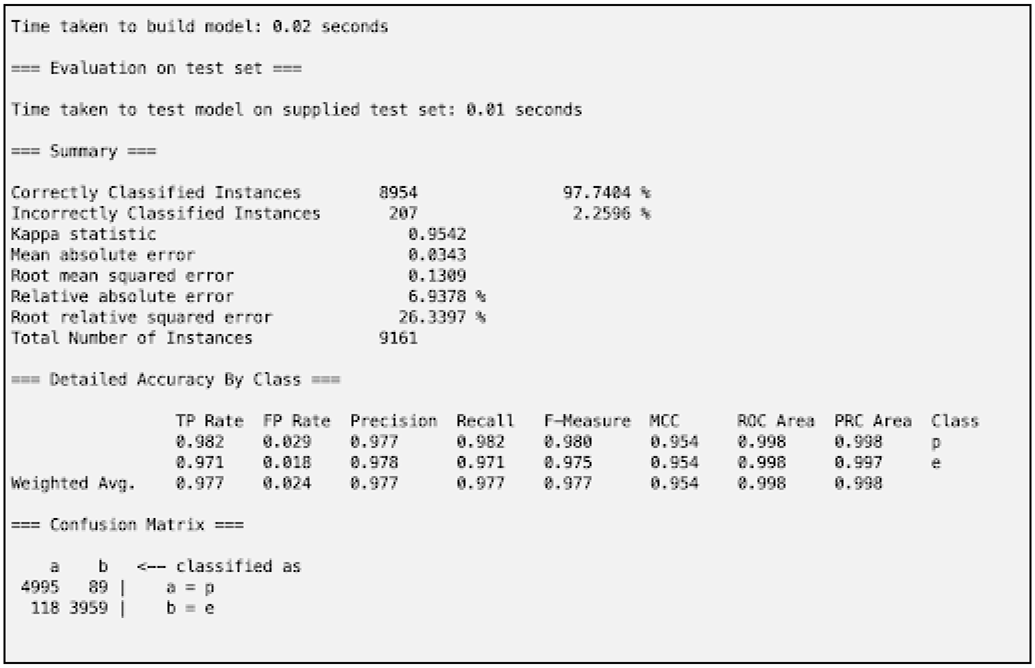
```
Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        9113               99.476  %
Incorrectly Classified Instances        48                0.524  %
Kappa statistic                          0.9894
Mean absolute error                      0.0097
Root mean squared error                  0.0698
Relative absolute error                  1.9737 %
Root relative squared error             14.049  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.007    0.995      0.996    0.995      0.989    0.999     0.999     p
                 0.993    0.004    0.995      0.993    0.994      0.989    0.999     0.999     e
Weighted Avg.    0.995    0.006    0.995      0.995    0.995      0.989    0.999     0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 5064   20 |    a = p
   28 4049 |    b = e
```

oner_test_set.arff, J48



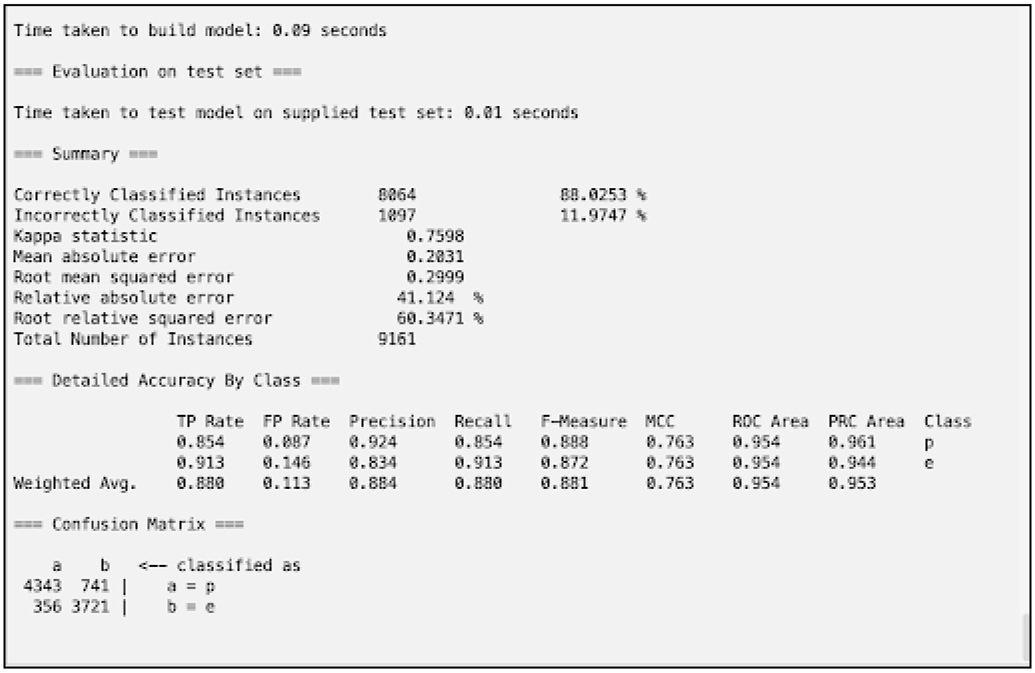
```
Time taken to build model: 0.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        9027               98.5373 %
Incorrectly Classified Instances       134                1.4627 %
Kappa statistic                          0.9704
Mean absolute error                      0.0561
Root mean squared error                  0.1381
Relative absolute error                 11.351  %
Root relative squared error             27.784  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.012    0.990      0.983    0.987      0.970    0.996     0.997     p
                 0.988    0.017    0.979      0.988    0.984      0.970    0.996     0.995     e
Weighted Avg.    0.985    0.014    0.985      0.985    0.985      0.970    0.996     0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 4999   85 |    a = p
   49 4028 |    b = e
```

selected_no_miss_test_set.arff, J48



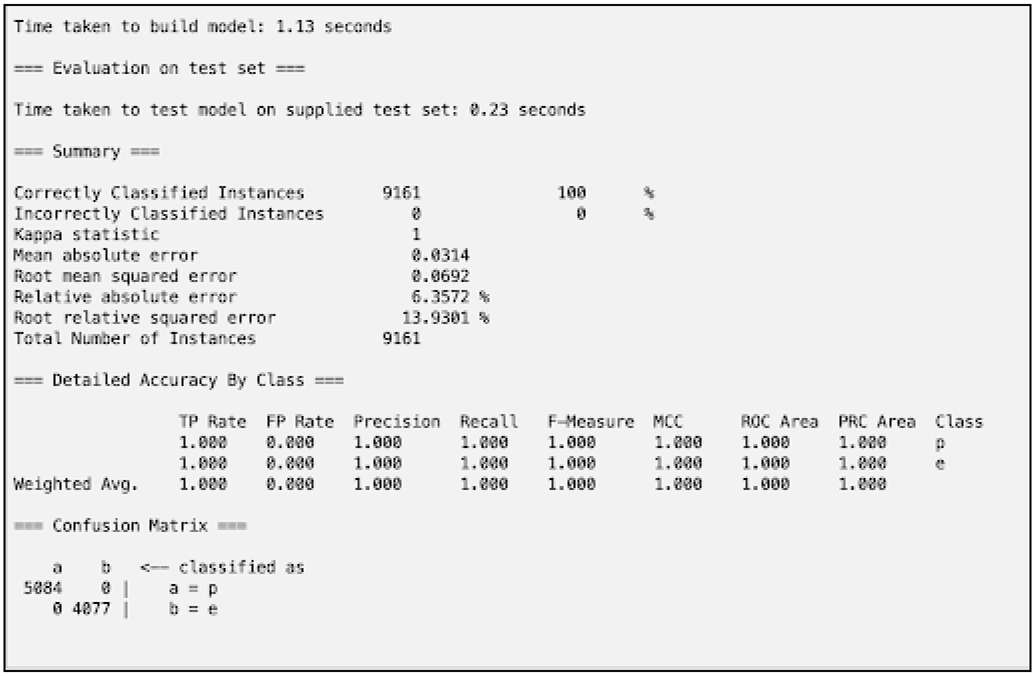
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8954               97.7404 %
Incorrectly Classified Instances       207                2.2596 %
Kappa statistic                          0.9542
Mean absolute error                      0.0343
Root mean squared error                  0.1309
Relative absolute error                  6.9378 %
Root relative squared error             26.3397 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.029    0.977      0.982    0.980      0.954    0.998     0.998     p
                 0.971    0.018    0.978      0.971    0.975      0.954    0.998     0.997     e
Weighted Avg.    0.977    0.024    0.977      0.977    0.977      0.954    0.998     0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4995   89 |    a = p
  118 3959 |    b = e
```

selected_test_set.arff, J48



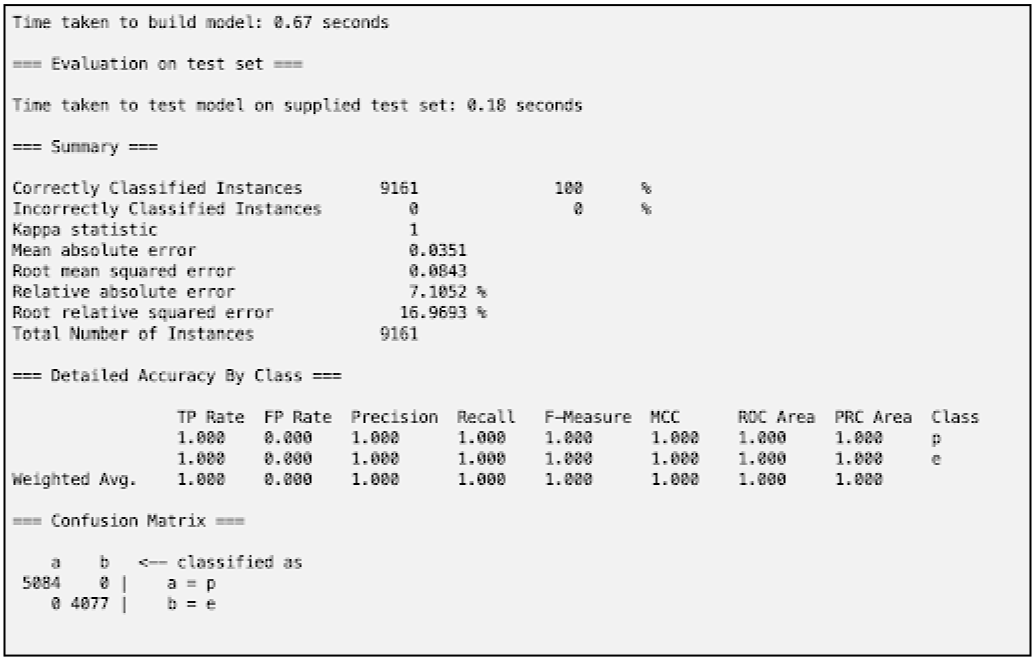
```
Time taken to build model: 0.09 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances        8064               88.0253 %
Incorrectly Classified Instances      1097               11.9747 %
Kappa statistic                          0.7598
Mean absolute error                      0.2031
Root mean squared error                  0.2999
Relative absolute error                 41.124  %
Root relative squared error             60.3471 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.854    0.087    0.924      0.854    0.888      0.763    0.954     0.961     p
                 0.913    0.146    0.834      0.913    0.872      0.763    0.954     0.944     e
Weighted Avg.    0.880    0.113    0.884      0.880    0.881      0.763    0.954     0.953

=== Confusion Matrix ===

    a    b   <-- classified as
 4343  741 |    a = p
  356 3721 |    b = e
```

cfs_test_set.arff, RandomForest



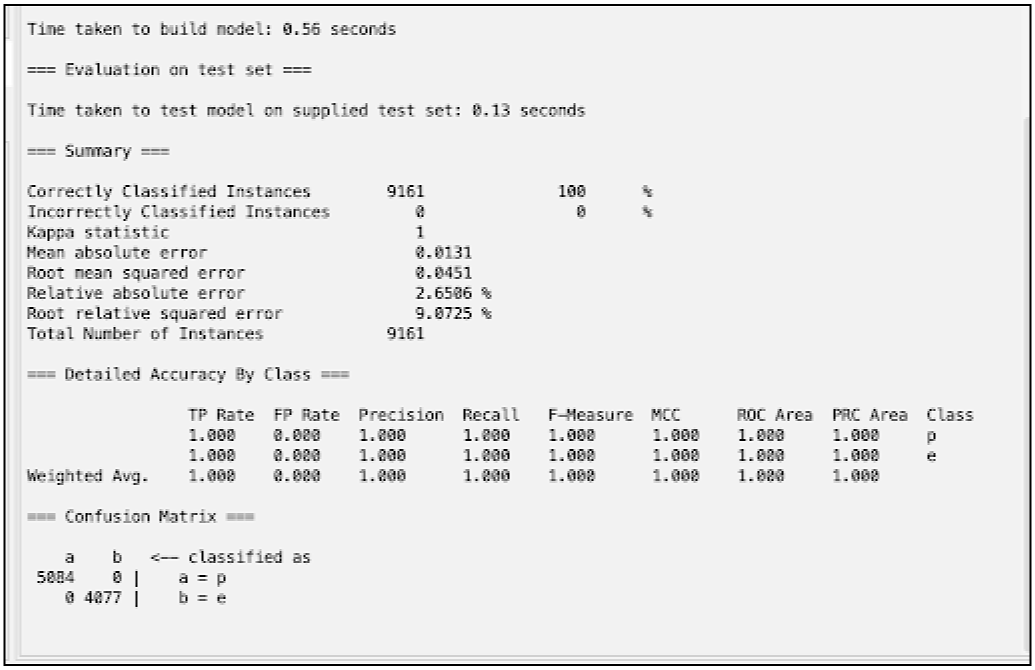
```
Time taken to build model: 1.13 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.23 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0314
Root mean squared error                  0.0692
Relative absolute error                  6.3572 %
Root relative squared error             13.9301 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

cfssubset_no_miss_test_set.arff, RandomForest



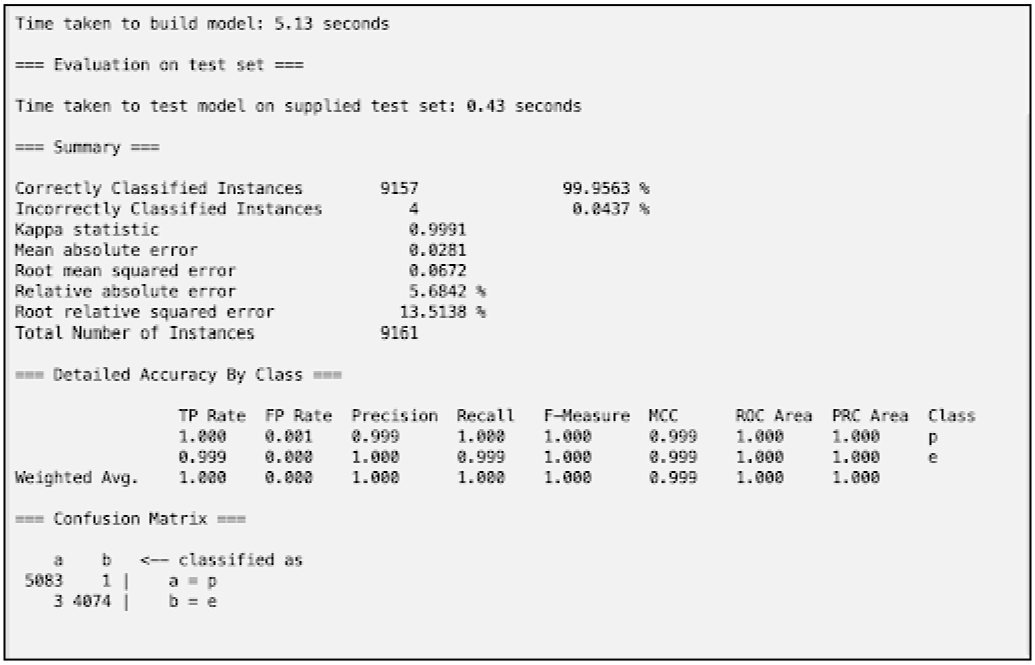
```
Time taken to build model: 0.67 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.18 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0351
Root mean squared error                  0.0843
Relative absolute error                  7.1052 %
Root relative squared error             16.9693 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

correlation_no_miss_test_set.arff, RandomForest



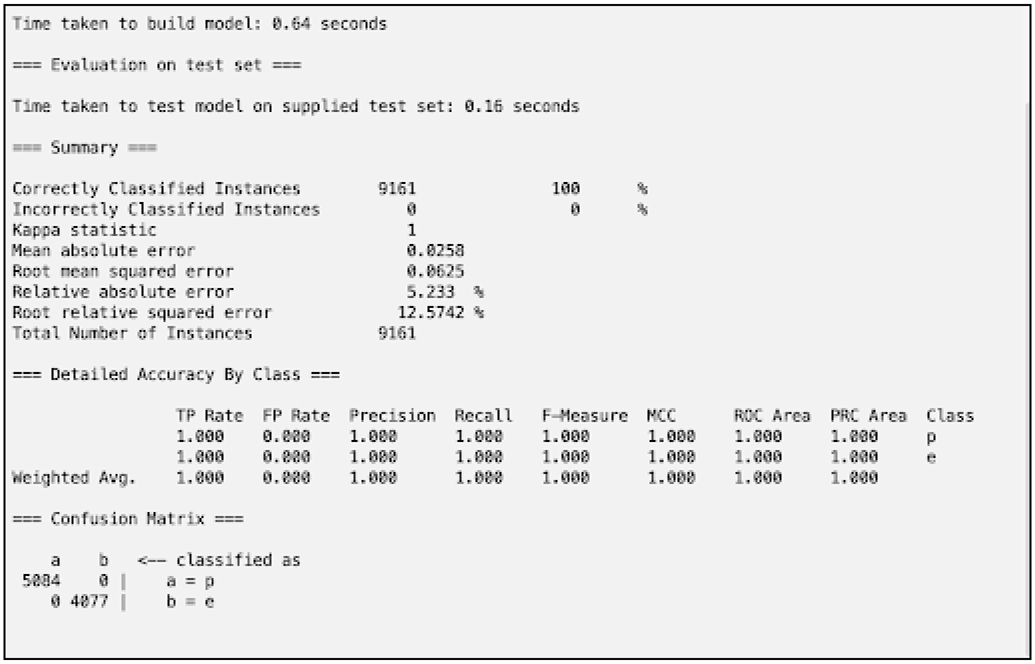
```
Time taken to build model: 0.56 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.13 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0131
Root mean squared error                  0.0451
Relative absolute error                  2.6506 %
Root relative squared error              9.0725 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

correlation_test_set.arff, RandomForest



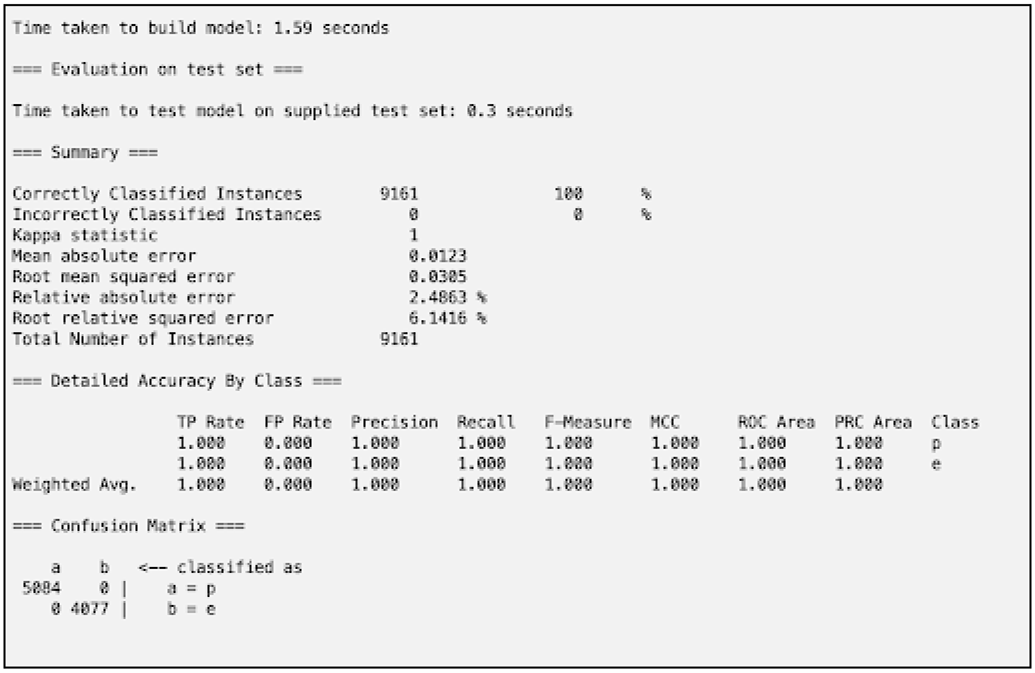
```
Time taken to build model: 5.13 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.43 seconds

=== Summary ===

Correctly Classified Instances        9157               99.9563 %
Incorrectly Classified Instances         4                0.0437 %
Kappa statistic                          0.9991
Mean absolute error                      0.0281
Root mean squared error                  0.0672
Relative absolute error                  5.6842 %
Root relative squared error             13.5138 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.001    0.999      1.000    1.000      0.999    1.000     1.000     p
                 0.999    0.000    1.000      0.999    1.000      0.999    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      0.999    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5083    1 |    a = p
    3 4074 |    b = e
```

infogain_no_miss_test_set.arff, RandomForest



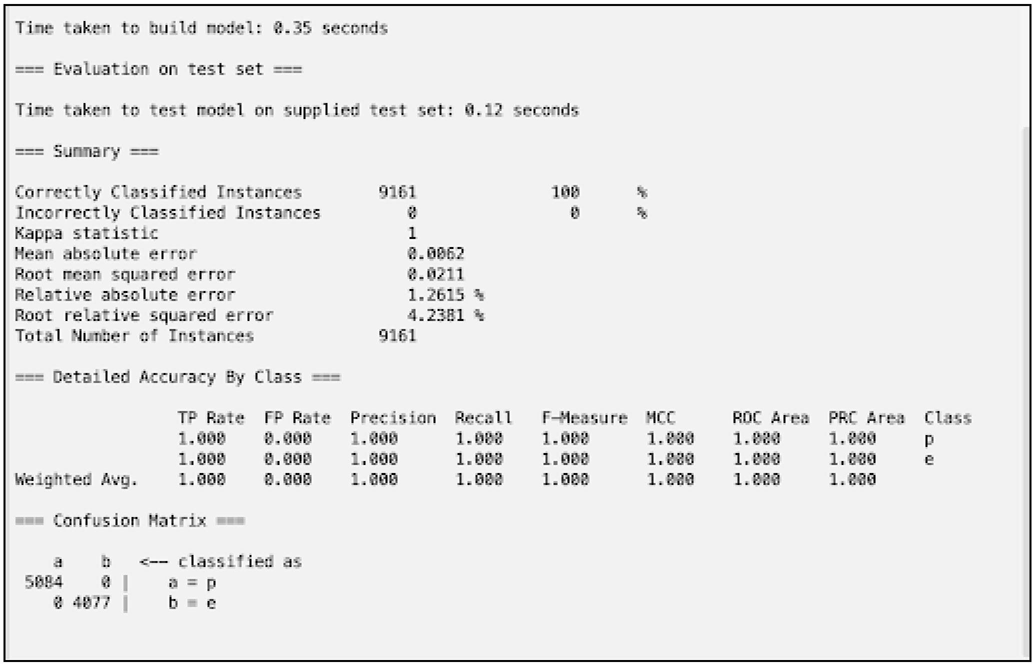
```
Time taken to build model: 0.64 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0258
Root mean squared error                  0.0625
Relative absolute error                  5.233  %
Root relative squared error             12.5742 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

infogain_test_set.arff, RandomForest



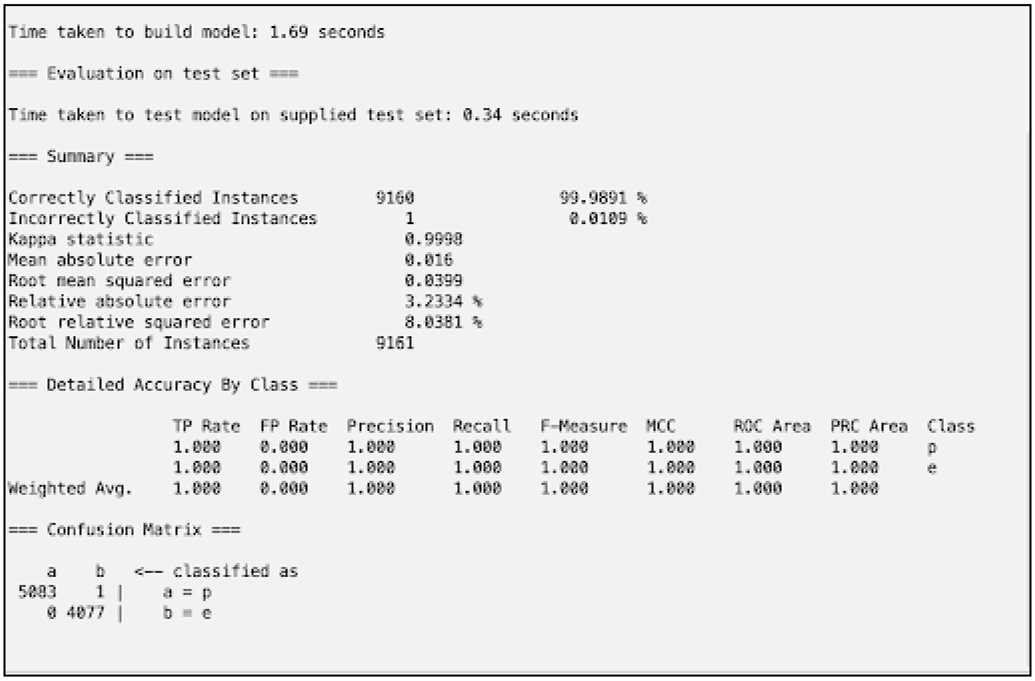
```
Time taken to build model: 1.59 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.3 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0123
Root mean squared error                  0.0305
Relative absolute error                  2.4863 %
Root relative squared error              6.1416 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

oner_no_miss_test_set.arff, RandomForest



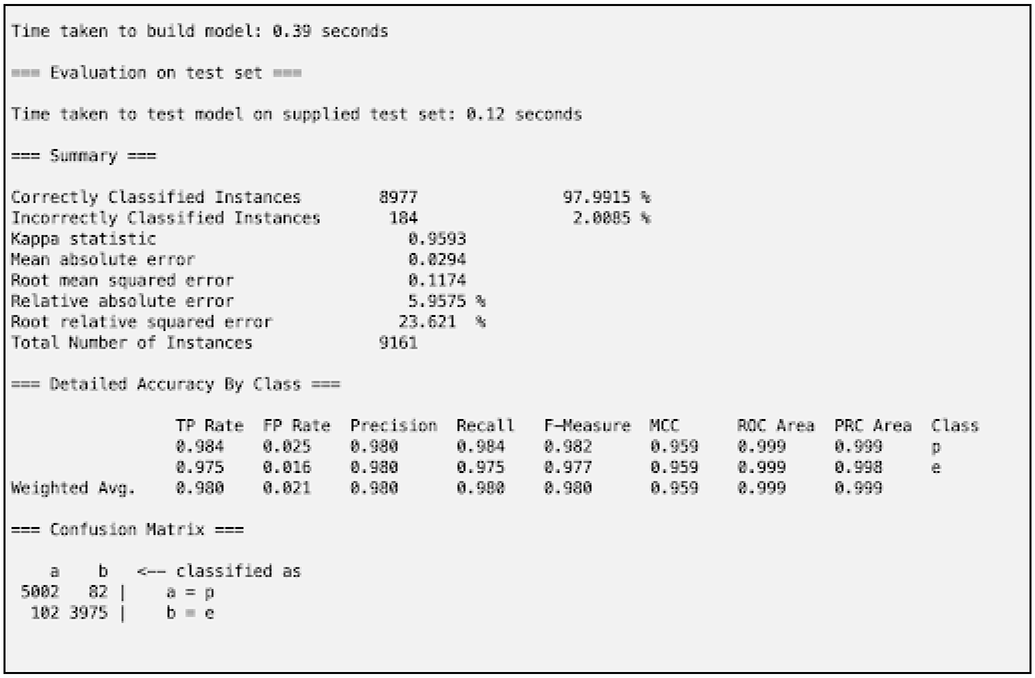
```
Time taken to build model: 0.35 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.12 seconds

=== Summary ===

Correctly Classified Instances        9161              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0.0062
Root mean squared error                  0.0211
Relative absolute error                  1.2615 %
Root relative squared error              4.2381 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5084    0 |    a = p
    0 4077 |    b = e
```

oner_test_set.arff, RandomForest



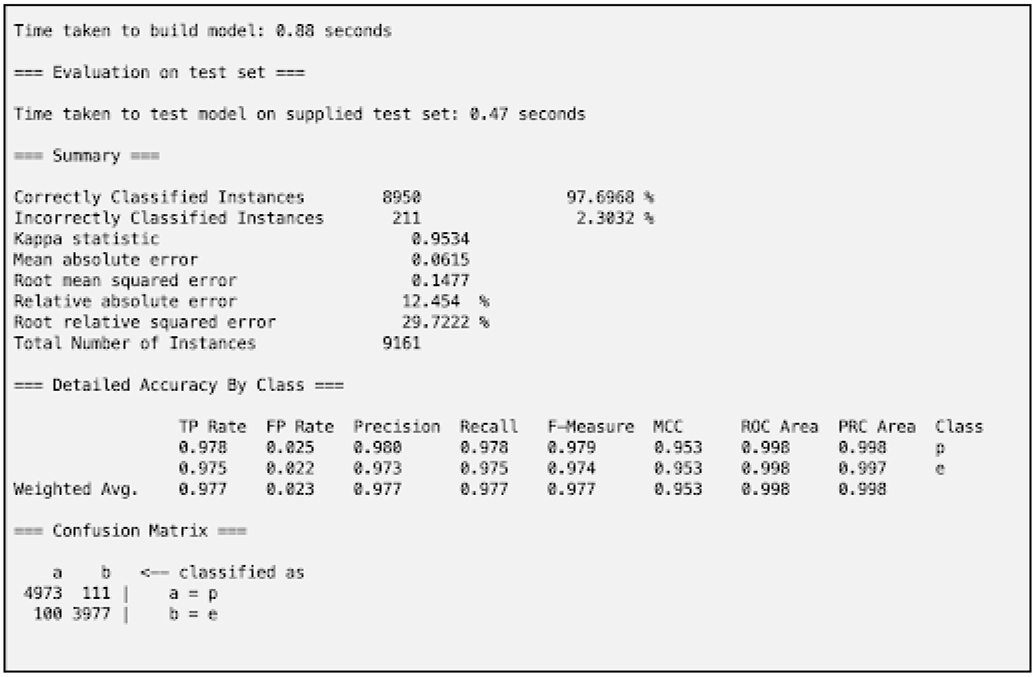
```
Time taken to build model: 1.69 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.34 seconds

=== Summary ===

Correctly Classified Instances        9160               99.9891 %
Incorrectly Classified Instances         1                0.0109 %
Kappa statistic                          0.9998
Mean absolute error                      0.016
Root mean squared error                  0.0399
Relative absolute error                  3.2334 %
Root relative squared error              8.0381 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 5083    1 |    a = p
    0 4077 |    b = e
```

selected_no_miss_test_set.arff, RandomForest



```
Time taken to build model: 0.39 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.12 seconds

=== Summary ===

Correctly Classified Instances        8977               97.9915 %
Incorrectly Classified Instances       184                2.0085 %
Kappa statistic                          0.9593
Mean absolute error                      0.0294
Root mean squared error                  0.1174
Relative absolute error                  5.9575 %
Root relative squared error             23.621  %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.984    0.025    0.980      0.984    0.982      0.959    0.999     0.999     p
                 0.975    0.016    0.980      0.975    0.977      0.959    0.999     0.998     e
Weighted Avg.    0.980    0.021    0.980      0.980    0.980      0.959    0.999     0.999

=== Confusion Matrix ===

    a    b   <-- classified as
 5002   82 |    a = p
  102 3975 |    b = e
```

selected_test_set.arff, RandomForest



```
Time taken to build model: 0.88 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.47 seconds

=== Summary ===

Correctly Classified Instances        8950               97.6968 %
Incorrectly Classified Instances       211                2.3032 %
Kappa statistic                          0.9534
Mean absolute error                      0.0615
Root mean squared error                  0.1477
Relative absolute error                 12.454  %
Root relative squared error             29.7222 %
Total Number of Instances             9161

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.978    0.025    0.980      0.978    0.979      0.953    0.998     0.998     p
                 0.975    0.022    0.973      0.975    0.974      0.953    0.998     0.997     e
Weighted Avg.    0.977    0.023    0.977      0.977    0.977      0.953    0.998     0.998

=== Confusion Matrix ===

    a    b   <-- classified as
 4973  111 |    a = p
  100 3977 |    b = e
```

Analysis

In total, we have 40 classifier models. Because we dealt with missing values two different ways, used 4 distinct classifiers - DecisionTable, NaiveBayes, J48, and RandomForest, and used 5 attribute selection algorithms (as outlined earlier in the report), we have ran a total of 40 models to classify the data. Among these:

cfs_test_set.arff, RandomForest
cfssubset_no_miss_test_set.arff, RandomForest
correlation_no_miss_test_set.arff, RandomForest
infogain_no_miss_test_set.arff, RandomForest
oner_no_miss_test_set.arff, RandomForest

All achieved perfect accuracies of 100%, as scored by WEKA. In order to select the best model among these 5, we must inspect their respective error loss metrics to determine which will perform best on real-life data sets. After further inspection, the RandomForest classifier with the OneR attribute selection algorithm and missing values filled with mean/mode yields a 0.0211 root-mean-square error, the lowest of the 5 aforementioned datasets. This means that even though our test set was large and thus more likely to be representative of real world samples if the data was taken correctly, it is even more likely that this model performs well.

Discussion & Conclusion

In this project, we learned how to use WEKA to successfully train and test a model that classified our own selected data of poisonous vs edible mushrooms with 100% accuracy. Our dataset initially started off with 20 attributes, and our model could have possibly been improved had we not removed some of the attributes during our attribute selection step. Many of the attributes that were removed could have held solid predictive power of whether or not a certain mushroom is edible or poisonous. In the future, better attribute selection could be a key improvement from our project.

Steps to Reproduce Our Model: RandomForest model with OneR Selection:

1. Open WEKA, click on Explorer, and load the secondary_data.csv file after clicking the Open file button.
2. Click on the Select Attributes button in the top, and under Attribute Evaluator, choose OneRAttributeEval and Ranker as the search method. Click start.
3. Keep note of the attributes which have a score less than 57.5, and remove them by returning to the Preprocess tab, checking their box, and clicking remove.
4. Save this dataset as a new file called oner_data.arff and input it into the split.py python script, returning 3 new files: oner_train_set.arff, oner_val_set.arff, and oner_test_set.arff.
5. Return to WEKA and click on the Classify tab on the top. Click Choose > trees > RandomForest as the classifier.
6. Click on the Supplied test set button and load in oner_test_set.arff.
7. Select the class on the left ((Nom) class).
8. Click start to test the model. Model has the name oner_model.model.

Team Contributions

Finding the Data & Building Proposal: Arya Bharath & James Wright
Preprocessing: Arya Bharath & James Wright
Intermediate Report: Arya Bharath & James Wright
Non-Weka Attribute Selection Algorithm: Arya Bharath & James Wright
Attribute Selection: James Wright
Classifiers: Arya Bharath
Results Analysis: Arya Bharath & James Wright
Final Report: Arya Bharath & James Wright

Appendix & Sources

Link to Dataset:
https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset

Attached Files:

- split.py - used to split each of our datasets into train, val, test sets.
- secondary_data.csv - original dataset.
- all .arff files - created from attribute selection and split, used to train and test model.
- oner_model.model - chosen model for this project.